Total Survey Error Summary

2018-2019 Fall

Total Survey Error and Data Quality – W1

References:

Introduction to survey errors

Groves, Robert, << Survey Errors and Survey Costs >>, Chapter 1 → Gbook

Biemer & Lyberg, 2003, << The survey process and data quality >>→ BL

Lecture notes → Lecture

Two kinds of **users** of survey data:

Describers—interested in discovering property of a fixed set of people. E.g., mean, proportion

Modelers—hypotheses about the case of social phenomena. E.g., associations

How to differentiate different **types of errors**:

Sampling errors—Errors as a result of drawing a sample rather than census

Non-sampling errors—Other component related to the data collection and processing procedures (e.g., non-response and measurement errors)

Another framework for differentiating different **types of errors**:

Error of observation—Error arising because of deficiencies in the measurement process. Causes further differentiates to different parties of the measurement process:

Instrument

Model of data collection

Interviewer

Respondent

Error of non-observation—Errors arising because measurements were not taken on part of the population. Further differentiates to:

Coverage error

Sampling error

Non-response error

Two kinds of research work **addressing errors**:

Measurer—focus on measuring the size the errors

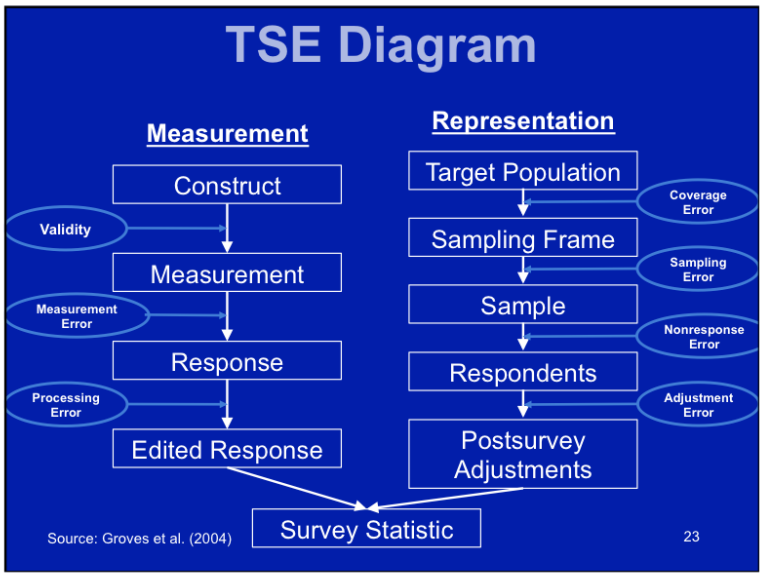
Reducer—focus on reducing errors

Survey process has two major lines: Development **measurement** and sampling **representation** of the population. Different errors occur along the steps (lecture):

Validity:
Extent to which measure reflects target construct

Measurement Error:
Extent to which response reflects "true" value.
-Interviewer
-Respondent
-Instrument
-Mode

Processing Error:
arises from processing activities that follow data collection (data entry, editing, coding)



Coverage Error:
Failure to include all elements of the target population in the sampling frame.

Sampling Error:
Failure to include all elements of the target population

Nonresponse Error:
Failure to measure all selected sample members

Adjustment Error:
Arise from postsurvey statistical adjustments (weighting, imputation, tabulation)

Key Concepts

Total Survey Error [TSE]

BL p34: includes components of errors that arise from sampling error and non-sampling error. It is the difference between a population parameter and the estimate of the parameter based on the sample survey.

Lecture: A conceptual framework used to broadly refer to the accumulation of all survey errors that may arise from the survey process.

Mean Square Error [MSE]

BL p45: One way to quantify the TSE for one specific survey estimate. The MSE gauges the magnitude of TSE on a particular estimate of interest. Unfortunately, MSE is not possible to compute directly from the survey data. Thus, this is also a theoretical concept.

MSE can break down to Variance and Bias:

$MSE = Variance + Bias^2$

Variance

Also named Variable Error

whether variance or bias is more harmful depends (BL p49)

for linear estimates (e.g., mean, proportion): bias is more damaging.

for non-linear estimates (e.g., correlation) variance weakens coefficients to zero.

The direction of bias's effect is uncertain

Gbook p9: Variable errors require an assumption of replicability of the survey. In principle, if the survey is replicated for many times, variable error is the part of the error is specific to each trail

BL p46: Errors that can be canceled out with each other if the survey is replicated for many times

Variance can be computed without knowing the true parameter value assuming replications of the survey

Question: For the conceptual replication, what are the components holding constant? Respondents? Interviewers? etc.

Bias

Also named Systematical Error

Gbook p8: The type of common to all implementations of a survey design.

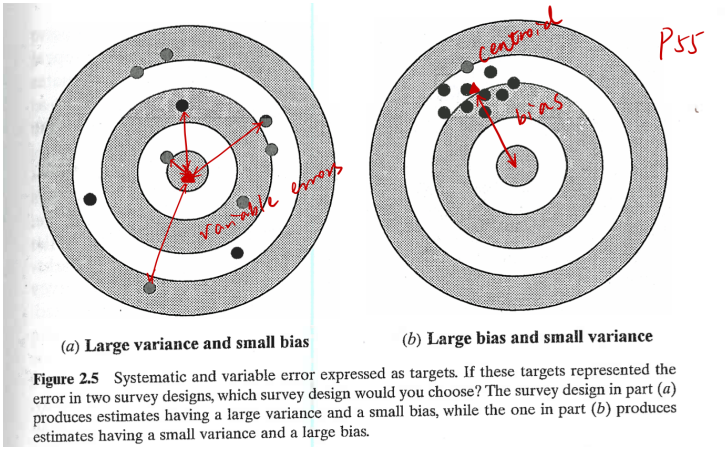
BL p47: Errors that cannot be canceled out. Errors that do not sum to zero when the sample observations are averaged.

Computing this requires knowledge of the true parameter. E.g., from administrative data.

The shooting a target figure in BL p 51 demonstrates the relationship and difference between variance and bias:

not specific to any specific survey in hand.

These are characteristics of the survey procedure.



(a) Large variance and small bias (b) Large bias and small variance

P55

Figure 2.5 Systematic and variable error expressed as targets. If these targets represented the error in two survey designs, which survey design would you choose? The survey design in part (*a*) produces estimates having a large variance and a small bias, while the one in part (*b*) produces estimates having a small variance and a large bias.

Two ways to compute MSE:

METHOD 1

MSE = average squared distance between the hits and the bull's-eye, or, in survey terms,
= average squared difference between the survey estimates from repeating the survey many times and the true population parameter value

METHOD 2

MSE = squared distance between the center of the hits and the bull's-eye + average squared distance between the hits and the center of the hits, or, in survey terms,
= squared distance between the average value of the estimator over replications of the survey process and the true population parameter value + average squared difference between the estimates from the replications and the average value of the replications

Accuracy

Gbook p15: the inverse of total error, including bias and variance. High accuracy = low bias + low variance

Precision

Gbook p16: mostly refer to the converse of variance. High precision = low variance

Standard Error

Gbook p16: square root of variance

Root Mean Square Error

Gbook p16: square root of mean square error (MSE) = square root of (variance + $bias^2$)

Measurement error

Gbook p16: depending on the specific definition, generally equates to observational error

Total Survey Error and Data Quality – W2 Cost + Coverage of the target population

Reference:

Groves, Robert. <<Survey Errors and Survey Costs>>, Chapter 2 + Chapter 3.1-3.5 → Gbook

Blumberg & Luke. 2009. Reevaluating the need for concern regarding noncoverage bias in landline surveys → Blumberg

Martin. 1999. Who knows who lives here?: Within-household disagreements as a source of survey coverage error. → Martin

Innaccione, Staab, & Redden. 2003. Evaluating the use of residential mailing addres in a metropolitan household survey. → Iannacchione

Lecture notes → Lecture

Cost

The **traditional** cost model is a linear model:

Total cost = Fixed costs + Variable costs

in which fixed cost is the cost to be incurred regardless of the sample size and variable costs vary as a function of design features.

For example, in the case of a stratification design. The variable cost of recruiting one case in each strata might differ, then the cost model can be specified as:

$$C = C_o + \sum_1^H C_h n_h,$$

where C_o = fixed cost, to be incurred regardless of what sample size is chosen;

C_h = cost of selecting, measuring, and processing each of the n_h sample cases in the hth stratum.

However, the traditional model does not fit well with the actual situation. The specification of the cost model should consider a variety of reality situations:

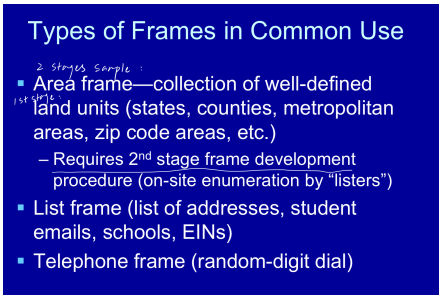
Nonlinear cost model often apply to practical survey administration (Gbook, p.57+Lecture):
Cost per case could decrease as the sample size increase because, for example, the interviewers learn and get increasing efficiently in handling each case.

Survey Cost models are inherently **discontinuous** (Gbook, p. 61+Lecture):
Add resources (recruiting a new interviewer or supervisor) every n cases

Cost models often have stochastic features (Gbook p. 63+ Lecture):
The costs vary because, for example, interviews have different efficiency and the difficulties of recruit different people are different. If the distribution of costs is highly skewed, then using the mean cost to specify the cost model would be problematic.

To tackle the problem of balancing error and cost in a survey design (Lecture):

1. Specify the error model as a function of the design features:
   In often cases, the most feasible error that can be specified in survey design is sampling error.
   Caveat: This can only be specified for one estimate/variable at a time
2. Specify the cost model
   Caveat: Most existing cost models do not account for complication characteristics of survey costs (e.g., nonlinearities, discontinuities and stochastic)
3. Combine the two models into optimization problem
   E.g., optimal sample size and allocation to strata that minimize error given a cost constraint



Frame and Frame Problem

What is "frame"? (Lecture)

Frame can be either physical lists (e.g., a list of email addresses) or procedures (e.g., random digit dialing). Frame can be used to (uniquely) identify, distinguish and access to the elements of the target population.

Gbook p. 100+ Lecture: Usually, a frame is available beforehand for the purpose of sampling. This is, however, not always the case. Sometimes the frame construction and the sampling steps are combined. For example, only an area frame (e.g., zip codes, blocks etc) is originally available for a survey that concerns household unit. In this case, the actual list of housing units is made after sampling areas in which the survey will be conducted.

(Gbook p82+Lecture)

Following the concept of frame. Frame population is the persons accessible through the frame.

Frame Population: A set of persons for whom some enumeration can be made prior to the selection of the survey sample

A frame population often try to correspond to a target population.

Target population: The finite population that the survey is attempting to characterize

A further categorization under frame population:

Survey population is the population of potential respondents, the people who are capable and willing to respond. Some people do not respond even if they are reached; these people are not survey population.

Frame = Survey population + Non-response

Coverage error is the mismatch between target population and frame population:

Target population
- specify population units : individual ? household ? business ?
- specify scope : non-institutionalized population ?
- specify period : 2019

Coverage error = Target population – Frame population

Question: Is coverage error basically coverage bias?

NO. Gbook P121. coverage error contains both variance and error P131. But it's true that it is most often conceptualize as a bias

Considering a linear statistics Y (e.g., mean), its value based on the target population is the weighted average of its value based on the frame population and the noncoverage population:

$$Y = \frac{N_c}{N} Y_c + \frac{N_{nc}}{N} Y_{nc},$$

(N_c: covered by frame; N_{nc}: not covered by frame; Y: target)

Thus, coverage error can be specified as (Gbook p. 85+Lecture):

$$\text{Bias} = Y - Y_c = (N_{nc} / N) \times (Y_c - Y_{nc})$$

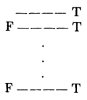
Recall that error specification is estimate/variable specific. So to make comparison across variables (which variable is subjected to more error?) ():

a. Rescale into relative terms: error/Y
b. Dichotomize variable (e.g., Blumberg Table 1): e.g.,
   Percentage of people who smoke. How biased is this percentage (+/- n%)?
   Percentage of people who drink. How biased is this percentage (+/- n%)?
   → Comparable

There are different ways from which a coverage error can arise:

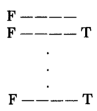
Under-coverage: Some elements of the target population do not appear in the frame population

Case II



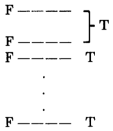
"Foreign" elements are included in the frame (Gbook)/ Over-coverage (Lecture): elements that are not of interest are included in the frame population. E.g., business numbers included in telephone survey of households

Case III



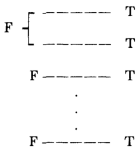
~~Over-coverage (Gbook)~~/ Multiplicity (Lecture): target population element has multiple appearances in the frame. These elements are more likely to be sampled.

Case IV



Clustering (Lecture): Mismatch between frame units and target units. Frame elements are household addresses and target elements are individuals.

Case V



Different kinds of coverage errors are not clear cut. For example, clustering error can be associated with under-coverage error because individuals within household can be omitted when households are contacted. This could due to confusion (not clear who "stay" in the household) or concealment (don't want to report the actual number of residents).

This point is illustrate by one of the reading of this week by Martin: 1999

There are errors and omissions in the process of constructing household rosters in household surveys. In complex and fluid households, who lives in the household is not clear. The unrelated people are more likely to be reported as living elsewhere, even if they themselves consider that household as their residence place. Similarly, the people who are often absent are likely to be considered as living elsewhere, even though they themselves do not think so.

Coverage error is unknown to researchers. But there are ways to estimate the error. Two of the three papers of this week illustrate the frame problem (coverage error) and potential ways to measure them:

-Secondary analysis of survey data where coverage (at least in principle is measured).
This is illustrate by Blumberg. This paper was interested in how bias landline surveys are. They rely secondary data of an in-person household survey, which is representative of the population. In this national survey, respondents were asked if adults live in landline household, cell phone only household, or no phone household. They then examined how biased landline people were, in comparison to the cell-phone people and no-phone people.

Blumberg & Luke, 2009

-Comparison with external benchmark.
E.g., Do the demographic composition of the current sample match with the census record?
Caveat: This does not only address coverage error. Mismatch between the current sample and external benchmark records is a result of aggregated errors (coverage+ non-response+ measurement errors).

-Demographic analysis.
A different way to count descriptive statistics.
E.g., Baseline population size+ morality rate, fertility rate, and immigration → Current population size
Caveat: This is useful for counting a population, but not so much if I want to know the characteristics of this population

-Post enumeration survey.
This is illustrate by Innacchione. They conducted on-site enumeration for a sample of residential

et. al. 2003

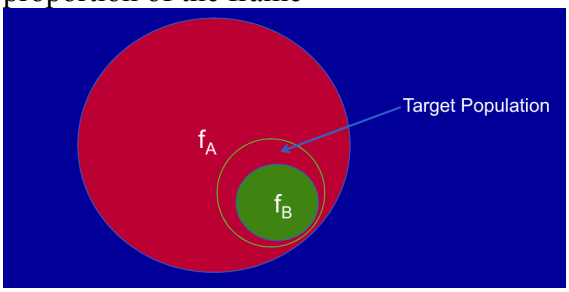
mailing lists to see 1) how many households are missed by the mailing list and 2) what the occupancy rate of the mailing list is.

Total Survey Error and Data Quality – W3
Repairs for Under-coverage

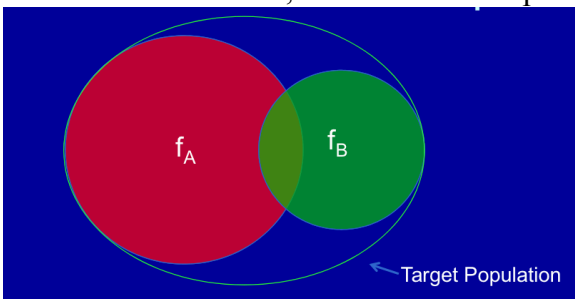
Reference:
Groves, Robert., <<Survey Errors and Survey Costs>>, Chapter 3 → Gbook
Iachan & Dennis. 1993. A multiple frame approach to sampling the homeless and transient population. → Iachan
Dever, Rafferty & Valliant. 2008. Internet surveys: can statistical adjustments eliminate coverage bias? → ~~Rafferty~~ Dever
Eckman, Does the inclusion of non-internet housholds in a web panel reduce coverage bias? → Eckman

Recall from last week that coverage error results from the mismatch between target population and the frame population.
To address this error:
1. Multiple frames
One single frame is not adequate, so use multiple frames:

- Frame not complete
- Or frame is complete, but inefficient because target population occupies a small proportion of the frame

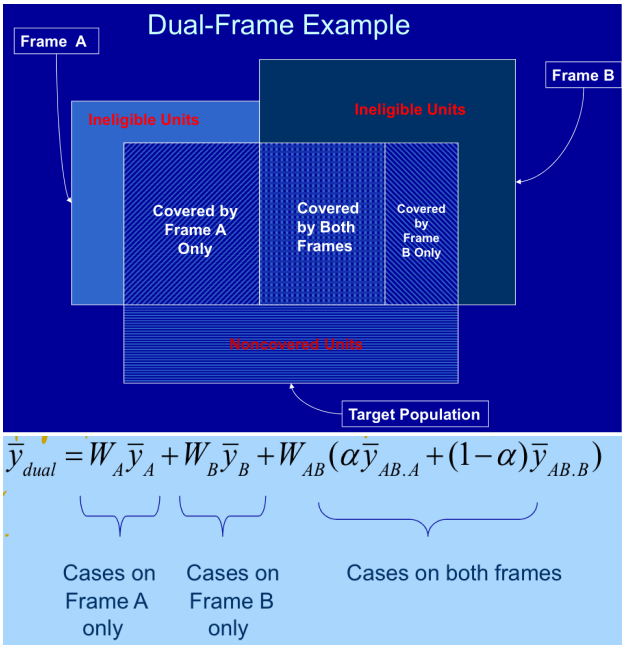


Frame A is inefficient; frame B is incomplete



Both frame A and frame B are incomplete.

In situation where multiple frames are involved, need to calculate the estimate in correspondence to the frame's relative size in the population.



$$\bar{y}_{dual} = W_A \bar{y}_A + W_B \bar{y}_B + W_{AB}(\alpha \bar{y}_{AB.A} + (1-\alpha)\bar{y}_{AB.B})$$

Cases on Frame A only | Cases on Frame B only | Cases on both frames

W_A, W_B, and W_{AB} are the porportions of the subgroup in the target population. α is a mixing parameter based on the proportion of the sample.

Multiple frames help to reduce non-coverage and reduce costs.
However, complication in using multiple frames involve:

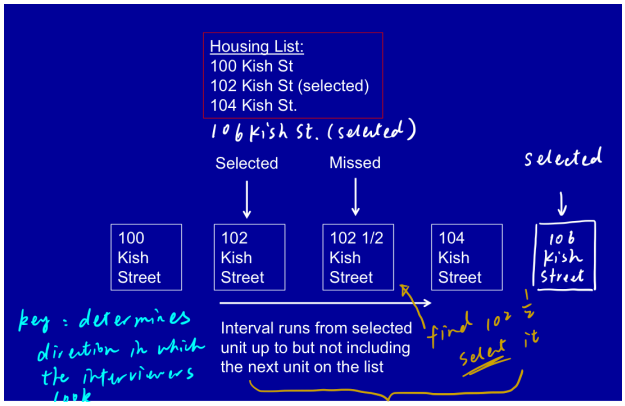
1 Determining frame membership. This is commonly achieved by asking each sampled unit if he/she also belongs to other frame.
2 Estimating the domain size (W_A, W_B, and W_{AB}).
3 Determining the optimal allocation between frames.
4 Integrating multiple frames to calculate sampling variance.

Ichan paper: An example demonstrating the useage of multiple frame in sampling the homeless and transient population. This paper exemplifies the four complication of integrating multiple frames.

2. Linking

List frames may not be perfect. Linking is basically a technique of frame construction. It updates the frame as the data collection goes. This method corrects under-coverage through adding missed units by linking them to the selected units.

Half-open interval (HOI) is one type of linking. It is commonly used for household survey. It links selected unit to all missed units that lie after it in the list frame, but before the next listed unit in the frame.



is a 104 is the next listed unit. Sample [102, 104) rather than [102]

This is an example. 102 ~~and 106 are~~ selected units. ~~Interviewers look for all missed units that lie after 102 but before 106.~~ If missed unit are identified, it is linked to the selected unit 102. The key is to determine the direction in which the interviewers look. While it seems intuitive on paper, it is not as easy in practice. A plan needs to be mapped out to tell the interviewers where to look.

Steps of HOI:

1) Build a frame in a pre-specified order
2) Define forward or backward links for checking missing units
3) Sample from the frame
4) For each sample unit, check for missing units according to linking rules

    e.g., Sample students from classes of a school. Attach potential new students to the last students on the list of each class. If the last student is selected, add the new students to the sample

Linking method helps to correct undercoverage error. However, complication in using HOI involves:

A) Requires the units in frame to be sorted in a meaning order
B) Ordinary linking procedures cannot deal with some type of missing units. For example, in household survey, if an apartment building is missing from the frame, it is hard to include all units in the apartment building in the sample, then a second-stage sampling is needed.
C) Increase uncertainty of sample size

Multiplicity Sampling is another example of linking method. = Network sampling, snowball sampling, respondent driven sampling. This method can be used when target population is hard to reach.

It utilizes the initial "frame" (i.e., seeds). During screening, informants report person in their networks who meet eligibility criteria. This method constructs the frame and conducts sampling at the same time.

This method reaches population that does not have a list frame to start from. However, complication involves:

a) makes strong assumptions about random selection. (Recently, more network analytical model helps to correct and control for covariates to meet this assumption)
b) Network members may not be locatable/accessible
c) Measurement error in linked to coverage error. Misreport/or refuse to report links would distort the frame.

3. Providing access to equipment to include population not included in a certain frame

Some units on a frame may not have the equipment required to participate in survey. For example, web survey would miss individuals without a computing device and internet access and people who don't have internet access are likely to be very different.

Providing access to equipment can remedy this problem.

Providing equipment not only address coverage issues, but can also tackle the measurement issues if equipment is provided to all respondents because the procedure is then standardized.

The advantage is that a) it reduces non-coverage, b) fast and cost-efficient web survey becomes possible, 3) can work with app-based data collection. However, complication of this method involves:

a) Lower recruitment rate

b) Expensive (more practical for panel studies)
c) Training and tech support/maintenance

Eckamn's paper uses data of LISS panel study. LISS put in much effort to recruit and retain non-internet household. Computer and internet were provided to the non-internet households if needed. This paper replicates five published studies which used LISS data and explored how the conclusion would have changed if the non-internet household were not included. While the internet household were demographically very different from the non-internet household, excluding these non-internet household did not have a large impact on the five studies' main conclusion. This study exemplifies that, at least in these five studies, under-coverage was not a serious issue for estimates of non-linear statistics (associations). Reparing for under-coverage did not make a significant contribution.

4. Statistical adjustment

Weighting by inverse of inclusion probabilities (base weight) does not combat coverage problems. It results in unbiased estimate of the frame, not the target population.
To correct for under-coverage, post-stratification is one method to go. This is a process that adjusts sample to known control totals. Gold standard of the population is needed. Then the sample is weighted up so that it becomes comparable to the population's demographic composition.
Post-stratification removes coverage errors if, within subgroups, expected values of covered are equivalent to non-covered. (Conditioning on the covariates that are used for weighting, the covered units are same as the uncovered unit)

Dever's paper is one study that exemplifies how weighting adjustment can repair the non-coverage error. BRFSS is a telephone survey that askes for respondents' internet access and health condition. The internet people did differ from the no-internet people on their health conditions. However, the difference can be largely attributed to differences in demographic covariates. As a result, weighting and adjusting the responses of the internet respondents based on covariates brought the estimates very close to the estimates based on both internet and no-internet respondents.

Benefit of weighting is that it virtually adjusts for non-coverage and non-response bias at the same time. However, complication involves:

a) Requires construction of subgroups and knowing subgroup sizes in the target population
b) Makes assumption that conditioning on subgroups, the covered and the non-covered are the same

One other "repair" for under-coverage is to redefine the target population

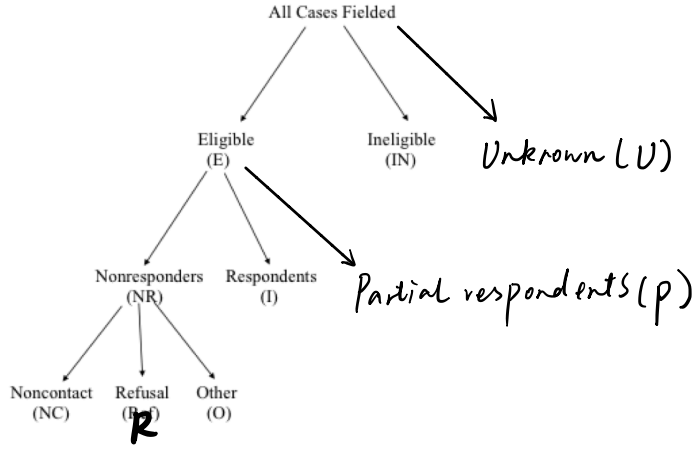
Total Survey Error and Data Quality – W4
Nonresponse Rates and Nonresponse Error

References
Groves and Couper. 1998. Nonresponse in household surveys. Chapter 1. → Groves&Couper
Curtin, Presser, & Singer. 2000. The effects of responde rate changes on the index of consumer sentiment. → Curtin
Groves & Peytcheva. 2008. The impact of nonresponse rates on nonresponse bias: A meta-analysis -→ Groves&PeytchevaMeta
Peytchev. 2013. Consequence of survey nonresponse. → Peytchev

1. Nonresponse Rates

Unit nonresponse is the failure to obtain survey measures on a sample unit. Unit nonresponse refers to the difference between the sample and the respondents.

Response rates—Proportion of *estimated eligible* sample who completed the survey. This rate is widely because it is easy and standardized to compute and one value applies to the whole survey.



$$RR1 = \frac{I}{(I+P)+(R+NC+O)+U}$$

$$RR2 = \frac{I+P}{(I+P)+(R+NC+O)+U}$$

$$RR3 = \frac{I}{(I+P)+(R+NC+O)+eU}$$

$$RR4 = \frac{I+P}{(I+P)+(R+NC+O)+eU}$$

eligible non response

estimated number of sample units of unknown eligibility that are eligible

Subjective part of this computation:
- Partial respondents—respondent or non-respondent?
- Eligibility unknown—denominator use U (count all unknown as eligible) or eU (Not all known are eligible → estimate the proportion of unknown that might be eligible)?

Other than response rate,
- can also calculate cooperation rate. Conditional on contact, the interviewed percent: $\frac{I}{(I+P)+R+O}$
- can also calculate refusal rate Percentage of eligible sample who refuse: $\frac{R}{(I+P)+(R+NC+O)+U}$

---

weighted response rate.

$$WRR1 = \frac{\sum w_i I_i}{\sum w_i (I_i + P_i + R_i + NC_i + O_i + U_i)}$$

---

measure of size weighted response rate.

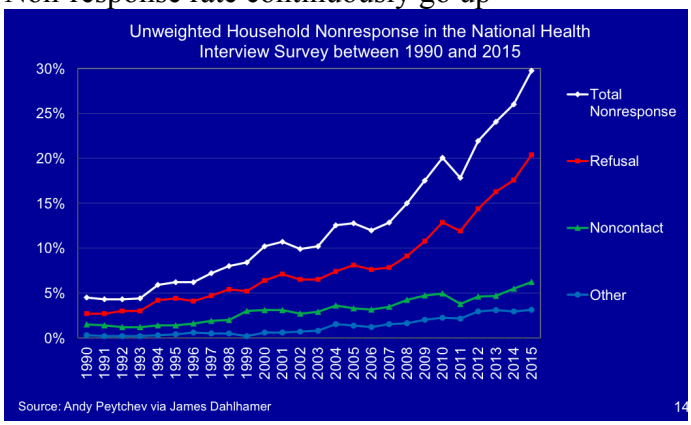
- more common for business survey. give a larger weight to larger business.

$$RRMOSW_1 = \frac{\sum w_i y_i I_i}{\sum w_i y_i (I_i + P_i + R_i + NC_i + O_i + U_i)}$$

where w_i is the inverse of selection probability and y_i is the measure of size.

2. Trends in nonresponse

Non-response rate continuously go up



This trend is observed across countries (West Europe and US) in different kinds of surveys. On average, 3% decline per year in cooperation rate (de Leeuw & de Heer, 2002).

Not only is the non-response rate increases, so is the effort required to recruit respondents. Costs have risen as survey takes countermeasures (advance letters, incentives, callbacks etc.).

3. Estimating the impact of nonresponse—Linking non-response rates to bias

Groves&Couper (p.11)+ Peytchev (p.90)+ lecture: Two views on nonresponse

Deterministic View: Each person in a target population either is a respondent or a nonrespondent for all possible surveys.

$$Bias = Y_r - Y = \frac{N_{nr}}{N}(Y_r - Y_{nr})$$

r and nr are fixed characteristics of population members. Bias depends on the proportion of the non-respondents, as well as the difference between the respondents and the non-respondents.

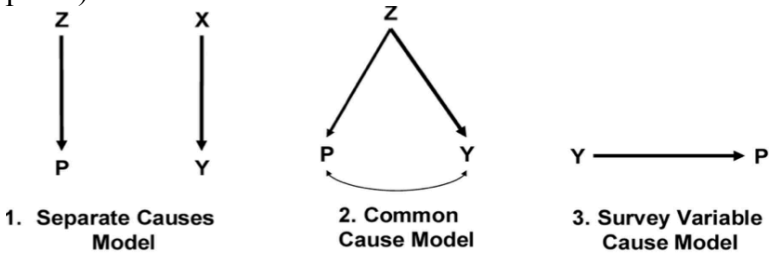
==========

Stochastic View: Each sample unit has a probability of being a respondent and a probability of being a non-respondent. Each sample survey is but one realization of survey design.

$$Bias(Y_r) = \frac{\sigma_{P,Y}}{\bar{P}}$$

where $\sigma_{P,Y}$ refers to the covariance between response propensity P and the survey variable Y, and $\bar{P}$ refers to the mean response propensity.

Below, the three graphs demonstrates three situations in which P and Y are independent, conditionally dependent, and dependent, respectively. Correspondingly, bias doesn't exist, can be controlled, and will always exist (Groves&PeytchevaMeta, p.168):



In situation *1 (Separate cases model)*, P and Y are independent, $\sigma_{P,Y} = 0$. The estimate of Y is unbiased. This corresponds to *missing completely at random (MCAR)*.
In situation *2 (Common cause model)*, P and Y are independent only if conditioning on Z (recall directed acyclic graphs). This corresponds to the *missing at random (MAR)* case. That is, there is a systematic pattern in missingness in Y, but this systematic pattern can be accounted for by observed data (i.e., by controlling for Z).

In situation *3 (Survey variable cause model)*, P is a caused by Y. There is no way to break the dependence between Y and P, so $\sigma_{P,Y} \neq 0$, and the estimate of Y is biased. This corresponds to *non-ignorable* condition of nonresponse.

==========

A cautionary note: Higher response rate can lead to higher non-response bias. Based on the *determinist view*, this means that while a procedure can push $\frac{N_{nr}}{N}$ to decrease, it results in the difference between respondents and non-respondents $(Y_r - Y_{nr})$ to increase. Based on the *stochastic view*, this means that while a procedure can pushes $\bar{P}$ to increase, it results in a stronger association between response propensity and variable of interest $(\sigma_{P,Y})$.

Groves&PeytchevaMeta paper demonstrates the lack of relationship between response rate and response bias. What is noteworthy of this figure is the large within study (same response rate) variability in nonresponse bias across estimates (yet, different response bias for different variables). Non-response bias (often the real interest) is a variable/estimate-level features. An overall non-response rate at the survey level doesn't necessarily capture what we care.

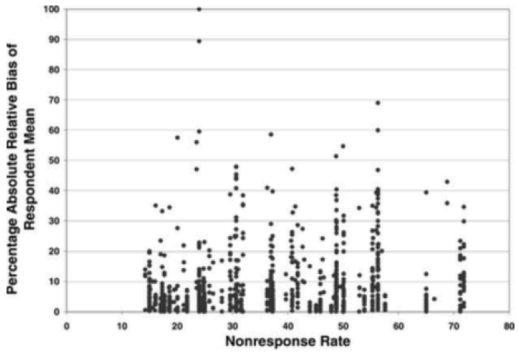


Figure 2. Percentage Absolute Relative Nonresponse Bias of 959 Respondent Means by Nonresponse Rate of the 59 Surveys in Which They Were Estimated.

Curtin is an empirical study that investigates the impact of non-response on estimates. Some respondents required much effort to recruit; if these efforts were not invested, they would become non-respondents. Curtin examined "what if these hard cases (cases that needed refusal conversion & cases than needed more than multiple calls) were not included?" They found no effect of excluding these respondents on estimate of ICS (index of consumer sentiment). This study exemplifies that non-response rate is not directly related to non-response bias.

Similar results demonstrated by Keeter (refer to lecture notes) who compared the response rate and bias of a standard survey and a rigorous survey: response rate differed? → Yes; response bias differed? → Not really

4. Causes of nonresponse

Noncontact
- Face-to-face: can't physically reach
- Telephone: don't pick up; block
- Web: spam filter

Non-cooperation:
- too busy
- self-absorbed & no community engagement (low social integration)
- barriers for unwanted intrusions

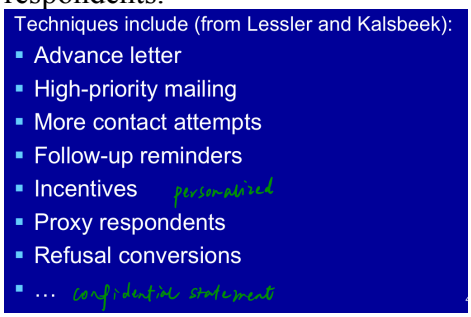
Inability to provide data:
- Physical/mental limitation
- Language barrier ← immigration

Total Survey Error and Data Quality – W5
Fixes for Nonresponse

References
Bethlehem. 2002. Weighting nonresponse adjustment based on auxiliary information. → Bethlehem
Lessler & Kalsbeek. 1992. Non-sampling errors in surveys. Chapter 8. → Lessler
Tourangeau, Brick & Li. 2017. Adaptive and responsive survey designs: a review and assessment. →Tourangeau
Groves. 2008. Issues facing the field: Alternative practical measures of representativeness of survey respondent pools. → Groves.etal

1. Prevention methods for unit nonresponse

Deal with non-response rate rather than the difference between respondents and non-respondents.



2. Identification of NR bias

The previous week demonstrates the lack of relationship between non-response rate and non-response. Going beyond non-response rate, other indicators are needed to indicate non-response bias (Groves.etal—entire article).

A single indicator at the survey level:

Often utilize *auxiliary information* (Bethlehem p.277): auxiliary information is defined as a set of variables that not only have been measured in the survey, but for which information on the sample or population distribution is also available.

a) Variance of non-response weight of *sample cases*
   = Variance of $\frac{1}{Response\ propensity}$
   Need auxiliary information on respondents and non-respondents to calculate response propensity
b) Variance of post-stratification weights of *respondents*.
   This is a bias from target population, rather than just a non-response bias.
c) Variance of response rates of subgroups.
   To calculate response rates of subgroups, the number of non-respondents in each subgroup is needed. Variance of response rate 1, 2, … j:

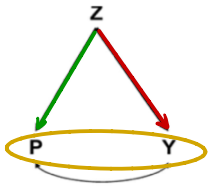
Subgroup 1 → response rate 1
Subgroup 2 → response rate 2
…
Subgroup j → response rate j

d) Goodness of fit statistics on propensity models
If goodness of fit statistics is good, it indicates that response is not random and can be well predicted. This suggests that there might be a non-response bias.

e) R-indicator
Again, dealing with response propensity by fitting propensity model using auxiliary data.
One version of r-indicator:
Response propensity: $0 \leq \rho \leq 1$ → Variance of ρ: $0 \leq S(\rho) \leq 0.5$
R-indicator: $0 \leq 1 - 2S(\rho) \leq 1$
When no variance in response propensity, $S(\rho)$=0, R-indicator = 1, which points to negligible NR bias. When $S(\rho)$=0.5, R-indicator= 0, which implies strong association between non-response and auxiliary variables.
There are different versions of R-indicators. Some are broken down based on each auxiliary variable.

Cautionary notes:
Essentially, these indicators are calculating variance in response propensity. This begs the question "what is the relationship between response propensity and response bias?":

Non-response bias ($\frac{\sigma_{P,Y}}{\bar{P}}$) is a function of response propensity ($\bar{P}$) and the covariance between response propensity and the variable of interest ($\sigma_{P,Y}$). Variance in P as a function of auxiliary variables (what these indicators are about) established the left arrow (marked green) of the below graph. This is a necessary but not sufficient condition for $\sigma_{P,Y}$. If there is indeed variance in P as a function of auxiliary variables (Z), then the covariance between P and Y is be established because we assume that there are also relationships between Y and auxiliary variables (Z):



2. Common Cause Model

→ This is what really needed for bias: σ_{PY}

But it is very possible that there is not relationship between auxiliary variables (Z) and variable of interest (Y) (right arrow marked red). Then variance in response propensity as a function of auxiliary variable tells nothing about the bias in Y.

Such discussion leads to another cautionary note. These indicators are survey level indicator. It is impossible for them to capture non-response error which functions at the estimate level. These indicators are used because it is assumed that in general, Z is associated Y(s).

Estimate level indicators:

a) Estimating the actual non-response bias on auxiliary information: comparing respondents and non-respondents on auxiliary variables

This basically is correlation between $\hat{P}$ropensity and Y. Exactly what we are interested in. The fact that $\hat{P}$ropensity is predicted by Z further points to our interest: correlation between Propensity and Y through Z.

1) If true P correlate with Y, but not through Z.
→ Z cannot predict P
→ $\hat{p}$ will not correlate with Y

2) If true P doesn't correlate with Y.
→ Say Z can perfectly predict P
→ $\hat{p}$ will not correlate with Y

3) If true P correlates with Y via Z
→ Z can predict P
→ $\hat{p}$ will correlate with Y.

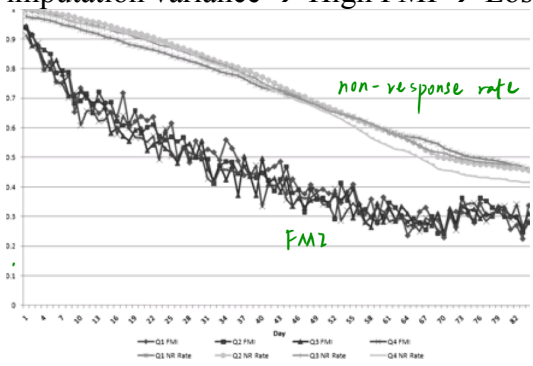
BAM ~

The amount of bias that we can control for.
Under MAC, it would be THE bias.

b) Correlation between post-survey adjustment weights and Y, on the respondent cases.
For example, if the adjustment weights are obtained by fitting a response propensity model based on auxiliary variables, then the adjustment weights are the summary of the relationship between Z and P (if there is no relationship, then all sample units would have the same predicted propensity and thus the same weight). Now we calculate the correlation between adjustment weights and Y, that links the left (green) and right (red) arrow of the above graph. A high correlation, would show a high dependence between P and Y through Z.

c) Examine the means of variable Y within deciles of the survey weight
Just a different version of the above method. Can be used to plot and generate visual display

d) Fraction of missing information (FMI) on Y
This is based on the ratio of between-imputation variance of an estimate and the total variance of an estimate, based on imputing values for all the non-respondent cases in a sample. When multiple imputation generates highly certain values, the between-imputation variance would be low, which would in turn mean that a lot of lost information (due to non-response) is recovered. FMI indicates the remaining lost information. (Highly uncertain imputation → High between-imputation variance → High FMI → Lost information remains high)



3. Compensations for NR bias

Adaptive and responsive survey design (Tourangeau)

(a) identifying a set of design features that might affect survey costs and errors,
(b) identifying indicators of the cost and error properties of those features,
(c) monitoring those indicators during the initial phases of data collection,
(d) changing design features (as needed) on the basis of those indicators and the cost–error trade-offs that they imply and
(e) combining the data from the various phases into a final data set.

Two common strategies:

1 Multiphase design—Initial phase informs data collection protocol in later phases. Mail → Telephone → Face-to-face

2 Case prioritization—Putting more effort to go after the low propensity and high value cases

An answered question (Tourangeau p. 217+ Lecture): The same variables that are available for propensity models (and thus guiding case prioritization) are also available for post-survey adjustment, so it is not clear whether balancing response propensity during data collection is necessarily more effective thatn simply adjusting the case weights afterward.

Double Sampling—Non-respondent follow up

Draw a subsample of non-respondents and use expensive protocol to measure the second phase sample of non-respondents. Estimate the characteristics of non-respondents to reduce non-response bias.

Complications of double sampling involve:

- Cost
- Added variance due to additional phase of sampling
- Hard to achieve 100% response rate in the second phase→ cannot fully reduce non-response error
- The success of the second phase is negatively correlated with the success of the first phase

Substitution (in the data collection stage)

Pull in new sample units to replace the non-respond sample units:

- Can either randomly substitute from the same subgroup
- Or rely on judgement to recruit a similar case

Advantage is that the final dataset should be balanced on auxiliary characteristics – no non-respondent. However, complications involve:

- Require knowledge about non-respondent (otherwise how to find substitute?)
- Is this still a probability sample? If random selection, then yes. If based on judgement, then no.

Weighting and Imputation

Important compensation of non-response—see next week

making a statement that "20% of the respondents drove impulsively"
This is just a description of the respondents. Without generalization, this is actually of no interest.

making a statement that "20% of 21-year old adults drove impulsively"
This is inference. This is of interest.

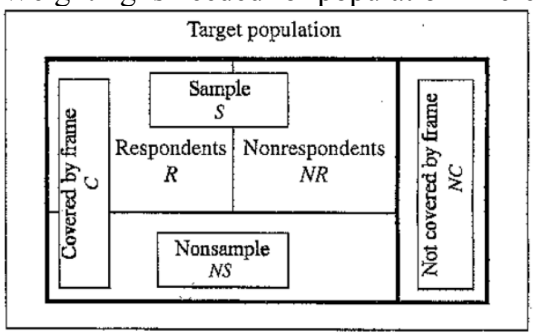
That's why we need to deal with non-response, to make inference.

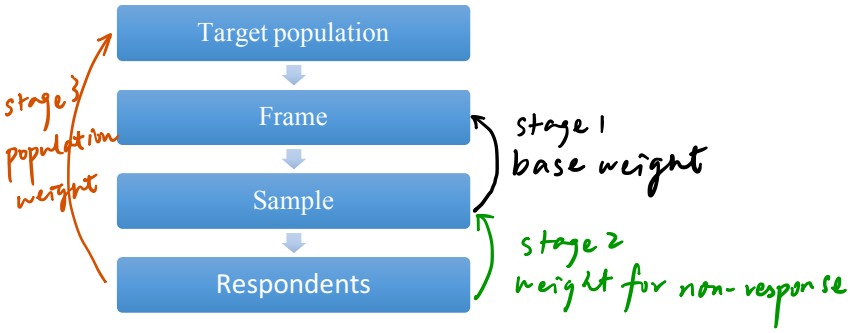
Total Survey Error and Data Quality – W6
Weighting and Imputation

Reference
Brick & Kalton. 1996. Handling missing data in survey research. → Brick&Kalton
Little and Vartivarian. 2005. Does weighting for nonresponse increase the variance of survey mean? → Little&Vartivarian
Rubin. 1986. Basic ideas of multiple imputation for nonresponse. → Rubin
Schenker et al. 2006. Multiple imputation of missing income data in the national health interview survey. → Schenker.etal

1. Weighting

Weighting is needed for population inference





Three stages of weighting (Brick&Kalton + Lecture):

a) First stage: Selection weight (base weight)

Account for unequal probabilities of selecting elements from the sampling frame
π_i=selection probability of case i
base weight of i = $w_{base,i} = \frac{1}{\pi_i}$

cv = standard deviation of weights / mean of weights

Variance of estimates typically increases with weights, as a function of the variability of the weight values across respondents. Proportional increase in variance (Little&Vartivarian, p.2+Lecture):
$L = cv^2$, where cv is the coefficient of variation of the respondent weights
Variance= (1+ L)*Variance

If simple random sampling, then w_i will the same for all sample units;
If some subgroups are oversampled, the w_i will weight down units of these subgroups

b) Second stage: Compensation for non-response

Auxiliary information on both respondents and non-respondents is needed. The assumption is that *conditioning on* auxiliary variables, the respondents are the non-respondents are the same. Thus, weight up the respondents to represent the non-respondents.

Typically, the first and second stage weighting are calculated independently. E.g.: propensity-based model doesn't incorporate selection weights
Some people argue otherwise, but independent first- and second- stage weighting is the common practice.

In previous week, non-response bias $Y_r - Y = \frac{N_{nr}}{N}(Y_r - Y_{nr})$

Now, non-response bias of adjusted mean can be calculated as a weighted sum of the non-response biases of each of the class h:

$$\sum W_h \frac{N_{nr,h}}{N_h}(Y_{r,h} - Y_{nr,h})$$

If the assumption is true (conditionally on h, no difference between respondents and non-respondents), the non-response bias of adjusted mean would be 0.

cell-level weights
⇕
case-level weights

$$\bar{y}_s = \sum^H W_h \bar{y}_{Rh}$$
$$= \sum^H \frac{n_h}{n} \bar{y}_{Rh}$$
$$= \sum^H \frac{n_h}{n} \frac{\sum^{r_h} y_i}{r_h}$$
$$= \frac{\sum^H \sum^{r_h} \frac{n_h}{r_h} y_i}{n}$$
$$= \frac{\sum^H \sum^{r_h} \frac{n_h}{r_h} y_i}{\sum^H n_h}$$
$$= \frac{\sum^H \sum^{r_h} \frac{n_h}{r_h} y_i}{\sum^H n_h \sum^{r_h} \frac{1}{r_h}}$$
$$= \frac{\sum^H \sum^{r_h} w_i y_i}{\sum^H \sum^{r_h} w_i}$$

where $w_i = \frac{n_h}{r_h}$ ← This is the inverse of response rate in stratum h

Weighting class adjustment

Partition the ***sample*** into weighting classes based on available auxiliary variables/cross-tabulate available auxiliary variables, e.g.:

	Old	Middle-age	Young
Male	Response rate	Response rate	Response rate
Female	Response rate	Response rate	Response rate

r_h refers to response rate in weighting class h

for respondent i who belong to weighting class h: $w_{class,i} = \frac{1}{r_h}$

Note that this weight only applies to respondents, but information on non-respondents is needed to calculate r_h

Drawback of this approach is that need to cross tabulate the auxiliary variables, so the auxiliary variable has to be categorical and cannot incorporate a large number of auxiliary variables. As an alternative:

Propensity-based models

Fitting a response propensity model based on auxiliary variable. This would give each respondent an estimated propensity of responding. Weight up the respondents who have a low response propensity

p_i is response propensity of respondent i

propensity-based weight $w_{propensity,i} = \frac{1}{p_i}$

Either use $w_{propensity,i}$ directly, or group respondents who have similar weight values and assign each group one weight

Drawback of weighting class adjustment is strong point of propensity-based model: auxiliary variables can be continuous; can include many auxiliary variables

Note that auxiliary variables are included in propensity model as covariates. If all interactions between these variables are specified, then propensity-based weighting would be the same as the weighting class adjustment.

Impact of weighting on bias and variance

Weighting's impact on bias and variance is estimate specific

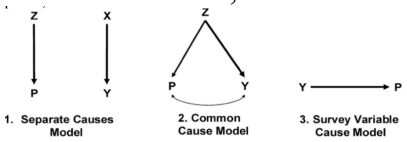
Little&Vartivarian paper discusses the impact of class weighting on bias and variance of mean estimate (p. 15)

Table 1. Effect of Weighting Adjustments on Bias and Variance of a Mean, by Strength of Association of the Adjustment Cell Variables with Nonresponse and Outcome.

	Association with outcome	
Association with nonresponse	**Low**	**High**
Low	Cell 1 Bias: ---- Var: ----	Cell 3 Bias: ---- Var: ↓
High	Cell 2 Bias: ---- Var: ↑	Cell 4 Bias: ↓ Var: ↓

Weighting can only remove bias when the auxiliary variable is highly associated with both response propensity and the variable of interest.

This makes sense, recall:



Bias exists when response propensity (P) is associated with the variable of interest (Y) (situation 2 and 3). Bias can be removed only when

1) P is not a direct causal effect of Y (i.e., not 3) &
2) Z which is strongly associated with both P and Y is identified and controlled for (i.e., only 2 and Z is conditioned on).

when choosing auxiliary variables, definitely go for the ones that correlates with the outcome variable.

As for variance, variance would decrease if the auxiliary variable based on which weighting is performed is strongly associated with the variable of interest. This is contrary to the common belief that weighting results in an increase in variance.

Taken together, ideally weighting should be conducted based on auxiliary variables that associates with both response propensity and variable of interest. But this ideal situation can hardly be achieved when weighting is at unit level.

c) Third stage: Account for under-coverage

should apply the sample weights in step 1 and 2 first.

The population weights built on top of the weighted sample.

Rely on external information on the target population total. To adjust for errors that derived from the current respondents being not representative of the target population, the population composition needs to be known.

Strictly speaking, this accounts for both under-coverage and non-response. It is an all-compassing method that bring the respondents composition back to the target population.

weighted ↓

Post-stratification

Partition the ***respondents*** *(contrast to weighting class adjustment, that's partition the sample)* into weighting classes based on population statistics

Ranking

	Male	Female
Respondents	30% n_{male}	70% n_{female}
Population	52% N_{male}	48% N_{female}

for respondent i who belong to weighting class j: $w_{population,i} = \frac{N_j}{n_j}$

e.g., weight for male respondent is $\frac{52\%}{30\%}$

Ranking

Useful when the external information is not enough to cross-tabulate. Iterative process. Adjust based on the first variable, then the second, the third … back to the first, until converge

Sum up the three stages of weighting:
Total weight = base weight * weight for non-response * population weight
$[w_{base,i} = \frac{1}{\pi_i}]$* $[w_{class,i} = \frac{1}{r_h}$ or $w_{propensity,i} = \frac{1}{p_i}]$*$[w_{population,i} = \frac{N_j}{n_j}]$

2. Imputations

Facing unit non-response, can:

- Do nothing
- Use prior wave's value
- Use sample mean
- Selection of nonmissing unit as "donor"
- Stratified selection of nonmissing unit as "donor"

Some common imputation method

- Hot deck
  - Order dataset, then select cases to impute (e.g., select the last case, select last case with criteria, random select with criteria)

- Regression (see the below equation)
  - Often cases, many variables have missing cases (i.e., multiple $\boldsymbol{y}$, all have missing)
    - **Sequential regression** (Lecture):
      Order variables based on level of missingness, work with the variables that have least missing first, then move up to work with variables that have more missing

    - **Iterative procedure** (Brick&Kalton, p.231):
      e.g., to impute for n variables: $\boldsymbol{y} = (y_1, y_2, \dots, y_n)$ using covariates $\boldsymbol{z}$
      First, a provisional imputation $\boldsymbol{z} \rightarrow y_1$;
      then $\boldsymbol{z} + \hat{y}_1 \rightarrow y_2$; …
      loop back... $\boldsymbol{z} + \hat{y}_2 + \hat{y}_2 + \cdots + \hat{y}_n \rightarrow y_1$ at least 5 circles → get rid of the order effect ↓ one imputation done

Brick&Kalton (p.227): one basic distinction between imputation methods is whether to set the error residual as 0. For example, consider a multiple regression imputation:

$$\hat{y}_{mi} = b_{r0} + \sum b_{rj} z_{mij} + \hat{e}_{mi}$$

where b_{r0} is the intercept and b_{rj} are the estimated regression coefficients for the regression of y on $\mathbf{z}$ obtained from the records with y values reported, z_{mij} is the value of z_j for record i with a missing y value, and $\hat{e}_{mi}$ is a residual term that is discussed further below. Most of the common imputation method can be represented by equation (3.1) with the appropriate definitions of $\mathbf{z}$ and $\hat{e}_{mi}$.

Deterministic approach $\widehat{e_{mi}} = 0$

- Single imputation

Stochastic approach $\widehat{e_{mi}} \neq 0$

- Multiple imputation under the stochastic approach

Not only e_{mi} is stochastic
$\hat{b}_{rj}$ can also be drawn from a distribution.

Most single imputations treat the dataset as a sample of n (sample units), rather than r (respondents), thus imputations result in an underestimate of variance because cases are reused.

Multiple ~~variance~~ imputation helps to restore the variance: use replication of the imputation to reflect the variance due to the imputation of data.

Impute multiple times and estimate the variance between these imputations

Variance estimate of multiple imputation is a function of *variance within imputation* and *variance between imputation*

The estimate is $\bar{\theta} = \frac{1}{m}\sum^{m} \hat{\theta}_m$

Let θ be the variable of interest to be imputed. A number of M imputations are conducted.

$$Var(\hat{\theta}_M) = \bar{U}_M + (1+\frac{1}{M})B_M$$

$\bar{U}_M = \frac{1}{m}\sum^{m} v(\hat{\theta}_m)$

where $\bar{U}_M$ is the mean within-imputation variance:

mean of (variance of $\hat{\theta}_1$, variance of $\hat{\theta}_2$, …,variance of $\hat{\theta}_M$)

where $B_M = \sum(\hat{\theta}_k - \hat{\theta}_M)^2/(M-1)$, with $\hat{\theta}_k$ being the overall mean of θ across all imputation and $\hat{\theta}_M$ being the mean of θ of each imputation

Schenker.etal paper is a empirical example of multiple imputation on a national survey dataset

$$\text{Total } v(\bar{\theta}) = \underbrace{\frac{1}{m}\sum^{m} v(\hat{\theta}_m)}_{\text{averaged within imputation sampling variance (of } \hat{\theta}_m)} + \frac{m+1}{m}\underbrace{\left[\frac{\sum_m (\hat{\theta}_m - \bar{\theta})^2}{m-1}\right]}_{\text{between imputation variance}}$$

↑ an extra factor

Elaboration on weighting

post-stratification

R's	Male	Female	Total
0-19	10	15	25
20-65	24	36	60
66+	5	10	15
Total	39	61	100

Pop'n	Male	Female	Total
0-19	15	14	29
20-65	29	30	59
66+	5	7	12
Total	49	51	100

Bring the (weighted) sample distribution to the population distribution.

r_h/r $\qquad$ N_h/N $\qquad$ $w_h = \dfrac{N_h/N}{r_h/r}$

weights at the element level $\Leftrightarrow$ weights at stratum level.

$$\bar{y}_{ps} = \sum_H w_h \bar{y}_h = \sum_H \frac{N_h}{N} \frac{\sum_{r_h} y_i}{r_h} = \frac{\sum_H \frac{N_h}{r_h} \sum_{r_h} y_i}{N} = \frac{\sum_H \frac{N_h}{r_h} \sum_{r_h} y_i}{\sum_H N_h} = \frac{\sum_H \frac{r}{N} \frac{N_h}{r_h} \sum_{r_h} y_i}{\sum_H \frac{r}{N} N_h}$$

$$= \frac{\sum_H \frac{N_h}{N} \frac{r}{r_h} \sum_{r_h} y_i}{\sum_H \frac{r}{N} N_h \left(\sum_{r_h} \frac{1}{r_h}\right)} = \frac{\sum_H \sum_{r_h} \frac{N_h}{N} \frac{r}{r_h} y_i}{\sum_H \sum_{r_h} \frac{N_h}{N} \frac{r}{r_h}} = \frac{\sum_H \sum_{r_h} w_i y_i}{\sum_H \sum_{r_h} w_i}$$

where $w_i = \dfrac{N_h/N}{r_h/r}$

Under the deterministic framework (refer to Elliott notes)

$$\text{Bias} = \bar{Y}_R - \bar{Y} = \underbrace{\sum_H W_h \frac{R_h - R}{R} (\bar{Y}_{hR} - \bar{Y}_R)} + \sum_H W_h (1 - R_h)(\bar{Y}_{hR} - \bar{Y}_{hNR})$$

post stratification gets rid of this part bc it uses respondents to represent the non-respondents and thus eliminates the effects of differences in response rates.

Under the stochastic framework:

$$\text{Bias}(\bar{y}_{ps}) = \sum_H W_h \, \text{Bias}(\bar{y}_{r,h}) = \sum_H W_h \frac{C_h(p, Y)}{\bar{P}_h}$$

Bias disappear if within strata, there is no covariance between response propensity p and variable of interest Y

Elaboration on weighting

Raking

match the marginal distributions of (weighted) sample to the marginal distributions of the population.

R's	Male	Female	Total
0-19	10	15	25
20-65	24	36	60
66+	5	10	15
Total	39	61	100

could be sample weighted already

Pop'n	Male	Female	Total
0-19	?	?	29
20-65	?	?	59
66+	?	?	12
Total	49	51	100

Final counts	Male	Female	Total
0-19	14.459	14.541	29
20-65	29.416	29.584	59
66+	5.126	6.874	12
Total	49	51	100

conceptually, raking is like imputation. We impute the distribution of cells that would allow us to satisfy all marginal distributions

The difference between post-stratification and raking is that raking doesn't provide information on interaction <u>at the population level</u>.

However, it also <u>doesn't enforce independence</u> at the population level. We don't impute $\frac{49}{100} \times \frac{29}{100} \times 100$ to '0-19 male', for example.

Instead, raking takes the middle ground. It preserve the dependence between variables <u>at the sample level</u>.

We don't have the interaction information at the population level to correct the sample distribution (we assume that the sample distribution is not good; that's exactly why we are weighting in the first place). Therefore, generally raking works better when we assume that there is no interaction between variables.

The stronger the interaction, the more the weights created by raking and post-stratification differ.

Raking estimator can lead to small cell and thus extreme weights.

$0.05 \times 0.05 \times 0.05 = 1.25 \times 10^{-4}$

None of it is too small by itself, but satisfying 3 marginal probabilities leads to a small cell.

Elaboration on imputation.

1. Mean value imputation.

Deterministic:

- replacing missing values with mean for the variable. Can distort the distribution, with spike at one value
- class mean imputation. impute $\bar{y}_k$

Stochastic:

$y_{ki} = \bar{y}_k + z_{ki}$ $\quad z_{ki}$ drawn from $N(0, \sigma_k^2)$

For binary variable, we can draw random number between [0, 1] and round with certain probability

2. Hot deck

Sequential hot deck imputation

The first value here is the mean

i	Gender	Educ	Reported family income	Hot Value	Imputation flag	Final value
1	M	9	23	*51*	0	23
4	M	11		*23*	1	23
2	M	12		*23*	1	23
3	M	12	43	*23*	0	43
7	M	12	35	*43*	0	35
8	M	12	42	*35*	0	42
5	M	16	75	*42*	0	75
6	M	16	88	*75*	0	88
16	F	10		*88*	1	88
15	F	12	28	*88*	0	28
17	F	12	31	*28*	0	31
18	F	12	35	*31*	0	35
19	F	12	30	*35*	0	30
22	F	12		*30*	1	30
13	F	14	67	*30*	0	67
14	F	15	56	*67*	0	56
21	F	15	72	*56*	0	72
20	F	18	66	*72*	0	66

- sort the list
- get value donnation from the one case above
- * can re-sort the list and donate again

multiple donation is a problem

boundary problem:
This boundary is bad.
The highest educated male donates his value to the lowest educated female.

A variation of sequential hot deck is hierarchical hot deck.

- group respondents and non-respondents into classes
- select donor at random within class.

↓

improve donnor-recipient match
reduce multiple donations

Gender	Education					
	<12		12		>12	
	R	M	R	M	R	M
Male	23	[...]	43 35 42	[...]	75 88	
Female		[...]	28 31 35	[...]	67 56 66	

no donnor → collapse classes

3. Regression imputation.

Missing on Y is to be imputated. Responses on X are complete.

Estimate the model $Y_i = \hat{\beta}_0 + X_{1i}\hat{\beta}_1 + X_{2i}\hat{\beta}_2 + \cdots + X_{pi}\hat{\beta}_p + \varepsilon_i$

Add the stochastic elements to both $\hat{\beta}$ and ε

$$\hat{\beta}_j \sim N(\hat{\beta}_j, \sigma^2_{\beta j})$$

$$\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$$

4. Sequential Regression imputation.

In the example above, $X_1 \cdots X_p$ are used to impute Y.

what if there are missing in $X_1 \cdots X_p$ too?

Refer to my summary 5-page above, under the "iterative procedure"

Total Survey Error
Measurement Error

Reference:
W8: Measurement Error Overview
Groves. 2004. Survey Errors and Survey Costs. Chapter 7. → Gbook
Biemer & Trewin. 1997. A review of measurement error effects on the analysis of survey data. → Biemer&Trewin
Fuller. 1987. Measurement Error Models. → Fuller

W9: Estimation of Measurement Error
Biemer & Stokes. 1991. Approahes to the modeling of measurement errors. → Biemer&Stokes
Saris, van Wijk, & Scherpenzeel. 1998. Validity and reliability of subjective social indicators: The effect of different measurements or assiciations. → Saris
Kreuter, Presser & Turangeau. 2008. Social desirability bias in CATI, IVR, and web surveys: The effects of mode and question sensitivity. → Kreuter
Antoun, Couper, & Conrad. 2017. Effects of mobile versus PC Web on survey response quality: A crossover experiment in a probability web panel. → Antoun

W10: Measurement Error: The Role of the Interviewers
Groves. 2004. Survey Errors and Survey Costs. Chapter 8. → Gbook
Hansen, Hurwitz, & Bershad. 1960. Measurement error in censuses and surveys. → didn't directly cite this paper, but the HHB model is discussed by several other readings
Conrad & Schober. 2000. Clarifying question meaning in a household telephone. → Conrad & Schober
West & Blom. 2016. Explaining interviewer effect: A research synthesis. → West&Blom
West & Olson. 2010. How much of interviewer variance is really nonresponse error variance? → West&Blom

W11: Measurement Error: The Respondents and the Questionnaires
Groves. 2004. Survey Errors and Survey Costs. Chapter 9-10 (didn't cite in this summary)
Krosnick & Presser. 2010. Question and questionnaire design. → Krosnick&Presser
Presser et al., 2004. Methods for testing and evaluating survey questions. → Presser.etal.
Gaskell, Wright, & O'Muircheartaigh. 2000. Teleschoping landmark events: Implications for survey research. → Gaskell.etal

Measurement error definition (Lecture7):
Two types of measurement error—Bias and variance

- Bias refers to systematic error, the reported values are consistently wrong
- Variance refers to random error, reported values are inconsistent/unreliable
- Thus, for a measure to be error-free, it must be consistent and free of bias
- Measurement error has meaning at both respondent and estimate level.
    - Consider $y_j = \mu_j + \varepsilon_j$
    - At the respondent level,
      Measurement bias means that the mean of ε_j's distribution ~~is~~ does center around not 0;
      Measurement variance means that ε_j has a distribution
    - At the estimate level,
      Measurement bias means that the realizations of ε_j for all the respondents do

not sum up to 0 (i.e., respondents' deviations do not average to 0); Measurement variance at the estimate level depends on how measurement error at the respondent level is conceptualized. If at the respondent level, ε_j is modeled as fixed (response deviations are fixed for each individual respondent/bias), the measurement variance at the estimate level is a function of the variation of respondent deviations (thus, if everyone deviates to the same extent in the same direction, no measurement variance). If at the respondent level, ε_j is modeled as variable over conceptual replications, the measurement variance at the estimate level is a function of the within-respondent variation (and between-respondent variation if there systematic respondent bias)

How to conceptualize measurement error in models:

Error model for ***continuous variable*:**

The basis and simplest situation simple random sampling without replacement (SRSWOR)—**Model 0** (Biemer&Stokes):

$$y_j = \mu_j + \varepsilon_j$$ where j refers to individuals

A serious of assumption:

1) Three sub-assumption about the true values

   a. $E(\mu_j) = \mu = \frac{1}{N}\sum_{i=1}^{N} \mu_i$

   b. $\mathrm{Var}(\mu_j) = \sigma_\mu^2 = \frac{1}{N}\sum_{i=1}^{N} (\mu_i - \mu)^2$

   c. $\mathrm{Cov}(\mu_j, \mu_{j'}) = -\frac{\sigma_\mu^2}{N-1}$, for $j \neq j'$ where N is the population size.

2) $E(\varepsilon_j | j) = 0$
   → For each individual unit (e.g., a respondent), the mean of its error is 0
   This assumption leads to $Cov(\mu_j, \varepsilon_{j'}) = 0$, for all j and j' → There is no covariance between any unit's true value and any error term (including itself).
3) $Var(\varepsilon_j | j) = \sigma_j^2$
   → For each individual unit, its error has a distribution
4) $Cov(\varepsilon_j, \varepsilon_{j'}) = 0$
   → Two units' errors are independent
5) Adding one component to the above setting:
   If the variable is measured α times (either through interview-reinterview, or through slightly different version of the question):

$$y_{j\alpha} = \mu_j + \varepsilon_{j\alpha}$$
$$Cov(\varepsilon_{j\alpha}, \varepsilon_{j\alpha'} | j) = 0$$

   → For the same individual, its error reported in each specific trail is random. Error in one trail is independent from error in another trail

There are two types of true value μ_j in the above equation.

- Platonic true score: Such measure has a *tangible true value* as an actual scalar score. For example, number of times I went to restaurant last week. There is a true score to this question. μ_j refers to this true score.
- Non-platonic/*Classical true score*: There is no single true score to such measure. For example, my attitudes toward premarital sex. "True value" is conceptualized as a distribution. μ_j refers to the *mean* of this distribution.

Relaxing different assumptions of Model 0 correspondents to different situations:

a) Relaxing assumption 2 $[E(\varepsilon_j|j) = 0]$ enables modeling *response bias*.

$$E(\varepsilon_j|j) \neq 0 \quad or$$

$$y_j = \mu_j + M_j + \varepsilon_j$$

where M_j is the "method effect" on this individual unit. For example, because of the wording of a question, unit j systematically underreports her age over all α times of the measurement.

b) Relaxing the deduction of assumption 2 $Cov(\mu_j, \varepsilon_{j'}) = 0$ enables modeling the association between response bias and true value. For example, people who have socially undesirable traits are more likely to underreport. When μ_j is undesirable, $E(\varepsilon_j)$ is negative

c) The above setting assumes that people are all subject to the same measurement operation. Different operations (e.g., different modes, different interviewers) can of course bring in additional errors and these errors will have a variation.

   i. Assigning different interviewers can be regarded as randomized assignment of measurement procedures to sample persons (Gbook p.304/305). Interviewer effect can be expressed as a violation of assumption 4 of Model 0 $[Cov(\varepsilon_j, \varepsilon_{j'}) = 0]$ (Biemer&stokes, p.493, **HHB Model**)
   Suppose that n units in the sample are collected by i interviewers, each interviewers work with m_i cases:

$$y_{ij\alpha} = \mu_{ij} + \varepsilon_{ij\alpha}$$

The assumption 4 is modified to:

4′. $\text{Cov}(\varepsilon_{ij\alpha}, \varepsilon_{i'j'\alpha}) = 0$ **for all** $i \neq i'$, different interviewers

$= \rho_w \sigma_\varepsilon^2$ **for all** $i = i', j \neq j'$ different units under same interviewer

for $-1 \leq \rho_w \leq 1$.

Then

$$\bar{y}_\alpha = \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{m_i} y_{ij\alpha}$$

Error terms of individuals within the same interviewer are correlated. But error terms of individuals within different interviewers are independent.

Under Model (0) with assumption 4′, Bias($\bar{y}_\alpha$) is still 0, but now

$$\mathrm{Var}(\bar{y}_\alpha) = \frac{1}{n^2}\Bigg[\sum_i\sum_j \mathrm{Var}(y_{ij\alpha}) + 2\sum_i\sum_{j<j'} \mathrm{Cov}(y_{ij\alpha}, y_{ij'\alpha}) + 2\sum_{i<i'}\sum_j\sum_{j'} \mathrm{Cov}(y_{ij\alpha}, y_{i'j'\alpha})\Bigg]$$

same interviewer, different units

different interviewer different units

$$= \left(1 - \frac{n}{N}\right)\frac{\sigma_\mu^2}{n} + \frac{\sigma_\varepsilon^2}{n} + \frac{1}{n^2}\left[\sum_{i=1}^{I} m_i(m_i - 1)\rho_w\sigma_\varepsilon^2\right]. \qquad (24.10)$$

The latter term is called the *correlated component of response variance* of the sample mean. A parameter which is sometimes used to describe the magnitude of the correlated component is the intra-interviewer correlation coefficient,

According to 4'

$$\rho_y = \frac{\mathrm{Cov}(y_{ij\alpha}, y_{ij'\alpha})}{\mathrm{Var}(y_{ij\alpha})} = \frac{\rho_w\sigma_\varepsilon^2}{\sigma_\mu^2 + \sigma_\varepsilon^2} = \rho_w(1 - R) \qquad (24.11)$$

ρ_y is not straightforward to estimate bc $\mathrm{cov}(\varepsilon_{ij\alpha}, \varepsilon_{ij'\alpha})$ isn't straightforward

$\rho_w = \frac{\mathrm{Cov}(\varepsilon_{ij\alpha}, \varepsilon_{ij'\alpha})}{\sigma_\varepsilon^2}$ ②

correlation between error terms with in same interviewer

$\frac{\mathrm{Cov}(\mu_{ij} + \varepsilon_{ij\alpha}, (\mu_{ij'} + \varepsilon_{ij'\alpha}))}{\mathrm{Var}(\mu_{ij} + \varepsilon_{ij\alpha})}$ (constant, constant) = ① + ②

$= \frac{\mathrm{Cov}(\varepsilon_{ij\alpha}, \varepsilon_{ij'\alpha})}{\sigma_\mu^2 + \sigma_\varepsilon^2}$ ①

where R is the reliability defined in (24.9). Under conditions described in the next section, ρ_y can be estimated from survey data. These estimates

where $R = \frac{\sigma_\mu^2}{\sigma_\mu^2 + \sigma_\varepsilon^2}$ (see below the deduction for model 0)

Notice the difference between ρ_y and ρ_w:

ρ_y depends on the correlation within interviewers based on the observed value $y_{ji\alpha}$

ρ_w depends on the correlation based on error terms $\varepsilon_{ji\alpha}$

For the (seemingly complicated) $Var(\overline{y_\alpha})$ equation above, if ignore the finite population correction and assuming that all interviewers work the same workload, then it can be reorganized to:

$$Var(\bar{y}_\alpha) = \frac{Var(y_{ji\alpha})}{n}[1 + (m - 1)\rho_y]$$

→ Equation of the ***design effect***!

The variance of the mean under the impact of interviewer effect is an inflated version of the variance under SRS ($\frac{Var(\bar{y}_{ji\alpha})}{n}$=Variance of mean under SRS)

ii. An alternative to note the interviewer effect (Biemer&Stokes, p.495; **ANOVA Model**):

$$y_{ji} = \mu_{ji} + b_i + e_{ji}$$

where e_{ji} has the same assumption as ε_{ji} in Model 0: $E(e_{ji}) = 0$. Factoring out b_i, different individuals' e_{ji} are uncorrelated to each other.

b_i is an *operator error* which is assumed to be the same for all units in the ith operator's assignment; e_{ji} is the unit-specific or *elementary error* due to the respondents (Biemer&Trewin, p.605)

b_i can be modeled as a *random* effect if there are many interviewers: Each interviewer i invokes a systematic bias on his/her respondents; pooling interviewers' biases together, then these biases cancel out each other. b_i has a distribution → random effect across interviewers

[Note 1. Notice how this equation differs from the equation in situation a]. Individual j is assigned to method i. And method i has the same effect on all individuals assigned to it, which is why noting the situation as $y_j = \mu_j + b_i + \varepsilon_j$ isn't as clear]

[Note 2. Recall multilevel model and marginal model:
The difference between the notation in i and ii exactly refers to the difference between marginal and multilevel model.]

If b_i is a random variable:

Then $\mathbf{Var}(y_{ij}) = \sigma_\mu^2 + \sigma_b^2 + \sigma_e^2$, **where** $\sigma_e^2 = \Sigma_i \Sigma_j \sigma_{ij}^2/n$ **and**

$$\mathbf{Cov}(y_{ij}, y_{i'j'}) = 0 \quad \text{if } i \neq i'$$

$$= \sigma_b^2 \quad \text{if } i = i', j \neq j'.$$

Then variance of the mean (sampling variance) is:

$$\mathrm{Var}(\bar{y}) = \frac{1}{n}(\sigma_\mu^2 + \sigma_b^2 + \sigma_e^2)[1 + (m-1)\rho_y]$$

$$= \frac{\mathrm{Var}(y_{ij})}{n}[1 + (m-1)\rho_y]$$

where

$$\rho_y = \frac{\sigma_b^2}{\sigma_\mu^2 + \sigma_b^2 + \sigma_e^2}.$$

Reliability is

$$R = \frac{\sigma_\mu^2}{\sigma_\mu^2 + \sigma_b^2 + \sigma_e^2}$$

Again, same as above. The intra-interviewer correlation coefficient is:

$$\rho_y = \rho_w(1 - R)$$

[Note. more on reliability and validity later]

There are **two perspectives** (Biemer&Stokes) to look at the $y_{j\alpha} = \mu_j + \varepsilon_{j\alpha}$

Sampling perspective:

- View survey response as a "two stage random sampling". In one trail,
- First, an individual is sampled from the finite population of individuals → μ_j
- Second, within this individual, an error is sampled from his/her infinite population of errors → $\varepsilon_{j\alpha}$
- Focus on how the measurement error impacts survey statistics

Psychometric perspective:

- Often work with multiple items (α) measuring the same construct
- Focus on the variance-covariance structure between different items that aim at the same construct or on the relationship between responses and errors

Indices of Measurement Errors (Biemer&Stokes)
Recall $y_{j\alpha} = \mu_j + \varepsilon_{j\alpha}$

Under the assumptions of Model 0:

$$\mathbf{Bias}(\bar{y}) = E(\bar{\mu}) - \mu = 0$$

$$\mathbf{Var}(\bar{y}) = \mathbf{Var}(\bar{\mu}) + \mathbf{Var}(\bar{\varepsilon})$$

$$= \left(1 - \frac{n-1}{N-1}\right)\frac{\sigma_\mu^2}{n} + \frac{\sigma_\varepsilon^2}{n}$$

[Note. μ is independent from ε.
Recall the "sampling perspective", individuals (μ_j) sampled from a finite population of people, errors ($\varepsilon_{j\alpha}$) sampled from an infinite population of errors. No finite population correction is needed for $Var(\bar{\varepsilon})$ because ε is sampled from an infinite population of errors]

where

$$\sigma_\varepsilon^2 = \frac{1}{N}\sum_{j=1}^{N}\sigma_j^2.$$

[Note. This is the average error variance of all individuals]

The existence of σ_ε^2 lowers the precision of $\bar{y}$ (higher $Var(\bar{y})$). The sampling perspective aims at reducing σ_ε^2

The psychometric perspective aims at finding measures of high **validity**—High correlation between observed value and true value ($y_{j\alpha}$ *and* μ_j):

$$\text{Theoretical Validity} = Corr(y_{j\alpha}, \mu_j)$$

$$= \frac{Cov(y_{j\alpha}, \mu_j)}{\sqrt{Var(y_{j\alpha})Var(\mu_j)}}$$

Theoretical validity reflects the degree to which the indicator measures the true score.

Reliability is defined as

$$R = 1 - \frac{Var(\varepsilon_{j\alpha})}{Var(y_{j\alpha})}$$

Under the strict assumptions of Model 0, with some math, the above equation becomes:

$$TV = \frac{Cov(y_{j\alpha}, \mu_j)}{\sqrt{Var(y_{j\alpha})\cdot Var(\mu_j)}}$$

$$Cov(y_{j\alpha}, \mu_j) = E((\mu_j + \varepsilon_{\alpha j})\cdot\mu_j) - E(\mu_j + \varepsilon_{\alpha j})\cdot E(\mu_j)$$

$$= E(\mu_j^2 + \varepsilon_{\alpha j}\cdot\mu_j) - E(\mu_j)^2 - E(\mu_j)\cdot E(\varepsilon_{\alpha j})$$

$$= E(\mu_j^2) + E(\varepsilon_{\alpha j}\cdot\mu_j) - E(\mu_j)^2 - E(\mu_j)\cdot E(\varepsilon_{\alpha j})$$

$$= E(\mu_j^2) - E(\mu_j)^2 \quad \leftarrow \text{bc under model 0 } \varepsilon_{\alpha j} \perp \mu_j$$

$$= Var(\mu_j)$$

$$Var(y_{j\alpha}) = Var(\mu_j + \varepsilon_{j\alpha}) = Var(\mu_j) + Var(\varepsilon_{j\alpha})$$

$$TV = \frac{Var(\mu_j)}{\sqrt{Var(\mu_j)}\sqrt{Var(\mu_j) + Var(\varepsilon_{j\alpha})}} = \frac{\sqrt{Var(\mu_j)}}{\sqrt{Var(\mu_j) + Var(\varepsilon_{j\alpha})}}$$

$$\text{Theoretical Validity} = \frac{\sigma_\mu}{\sqrt{\sigma_\mu^2 + \sigma_\varepsilon^2}}$$

$$\text{Reliability} = 1 - \frac{\sigma_\varepsilon^2}{\sigma_\mu^2 + \sigma_\varepsilon^2} = \frac{\sigma_\mu^2}{\sigma_\mu^2 + \sigma_\varepsilon^2}$$

And thus, $Theoretical\ Validity^2 =$ Reliability

Bias.
There are different ways to operationalize bias. Assuming that we have a flawed measure Y and an error-free measure X (μ):

Individual Measures	Aggregate Measure
$\lvert Y_i - X_i \rvert$ Absolute difference	$\frac{\sum \lvert Y_i - X_i \rvert}{n}$
$(Y_i - X_i)^2$ Squared difference	$\frac{\sum (Y_i - X_i)^2}{n}$
$(Y_i - X_i)$ Signed difference	$\frac{\sum (Y_i - X_i)}{n} = \bar{Y} - \bar{X}$ Estimated bias
$\frac{Y_i - X_i}{X_i}$ Relative difference	$\frac{\bar{Y} - \bar{X}}{\bar{X}}$ Estimated relative bias

Error model for ***binary variable***:

All the above discussion applies only to continuous variable. If a variable is binary, error can only take two forms—the true value is 1 but the response is 0 (i.e., false negative); the true value is 0 but the response is 1 (i.e., false positive).

For individual j within ith operator assignment:

False negative: $\theta_i = P(y_{ij} = 0 | \mu_{ij} = 1)$
False positive: $\phi_i = P(y_{ij} = 1 | \mu_{ij} = 0)$

Thus, for the (familiar) model:

$$y_{ij} = \mu_{ij} + \varepsilon_{ij}$$

The distribution of ε_{ij} is:

$\boldsymbol{y_{ij}}$	$\boldsymbol{\mu_{ij}}$	$\boldsymbol{\varepsilon_{ij}}$	$\boldsymbol{P(y_{ij} \mid \mu_{ij})}$
1	1	**0**	$\boldsymbol{1 - \theta_i}$
0	1	**-1**	$\boldsymbol{\theta_i}$
1	0	**1**	$\boldsymbol{\phi_i}$
0	0	**0**	$\boldsymbol{1 - \phi_i}$

Assumption 2 that $E(\varepsilon_j | j) = 0$ has to be relaxed. Instead

At the respondent level:
$E(\varepsilon_j | j) = -\mu_j(\theta_i) + (1 - \mu_j)\phi_i$
Thus for the people whose $\mu_j = 1$, their $E(\varepsilon_j | j) = -\theta_i$;
for the people whose $\mu_j = 0$, their $E(\varepsilon_j | j) = -\phi_i$

At the estimate level for $\bar{y}$:

$$Bias = E_n\left(E\left(\varepsilon_j \middle| j\right)\right) = E_n\left(-\mu_j(\theta_i) + \left(1-\mu_j\right)\phi_i\right) = -\mu\theta_i + (1-\mu)\phi_i$$

In general, there are different ways to operationalize bias for a categorical variable:

		Survey Measure	
		0	1
True Score	0	a	b
	1	c	d

Measure	Formula
Gross discrepancy rate	$\frac{b+c}{n}$
Net discrepancy rate (Can false positive and false negative cancels out?)	$\frac{b-c}{n}$
False negative rate (miss rate)	$\frac{c}{c+d}$
False positive rate	$\frac{b}{a+b}$

Measurement Errors in **Associations** (Biemer&Trewin; Fuller)
The above discussion concerns how measurement errors make the observed values deviate from the true value. Such deviations would influence the mean and total estimate of this variable. Extending beyond this variable itself, how does measurement errors of a variable influence the association of this variable with other variables?

Regression independent variable x on dependent variable y

1) Consider the *simplest situation*:
   Only x *is subjected to measurement error*; the true value x^*, so:
   $x_{ij} = x_{ij}^* + \varepsilon_{ij}$
   But y is not subjected to measurement error

   Regression model:
   True coefficient: $y_{ij} = \beta_0^* + \beta_1^* x_{ij}^* + \eta_{ij}$
   Observed coefficient: $y_{ij} = \beta_0 + \beta_1 x_{ij} + e_{ij}$
   β_1 vs. β_1^*?

   It is easily provable that *the observed coefficient (β_1) is smaller than the true coefficient (β_1^*)* if independent variable x contains measurement error:

$x_i = x_i^* + \eta_i$

x_i: observed value; x_i^*: true value.

The true relationship between X and Y:

$$y_i = \beta^* \cdot x_i^* + \varepsilon_i$$

$$\beta^* = \frac{Cov(y_i, x_i^*)}{Var(x_i^*)} \Rightarrow Cov(y_i, x_i^*) = \beta^* \cdot Var(x_i^*)$$

The observed relationship between X and Y:

$$y_i = \beta \cdot x_i + \varepsilon_i$$

$$\beta = \frac{Cov(y_i, x_i)}{Var(x_i)}$$

$$= \frac{Cov(y_i, x_i^* + \eta_i)}{Var(x_i^* + \eta_i)}$$

$$= \frac{Cov(y_i, x_i^*) + Cov(y_i, \eta_i)}{Var(x_i^*) + Var(\eta_i)} \quad (Cov(y_i, \eta_i) \to 0)$$

$$= \frac{Cov(y_i, x_i^*)}{Var(x_i^*) + Var(\eta_i)}$$

$$= \frac{\beta^* Var(x_i^*)}{Var(x_i^*) + Var(\eta_i)}$$

$$= \frac{\sigma^2_{x_i^*}}{\sigma^2_{x_i^*} + \sigma^2_{\eta_i}} \cdot \beta^*$$

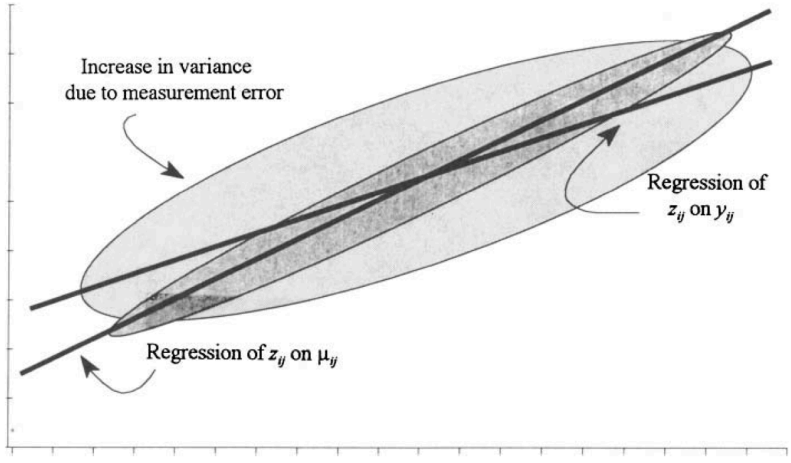
Cov(a+b, c) = Cov(a, c) + Cov(b, c)

η_i is independent ε_i ∴ $Cov(y_i, \eta_i) = 0$.

$\beta \leq \beta^*$ the observed coefficient is attenuated in comparison to true coefficient

Alternatively, an intuitive way of showing why the coefficient is smaller when there is measurement error in the independent variable x (Biemer&Trewin, p.619):

The error cloud *stretch horizontally* since now the x (x-axis) has measurement errors



2) Building on top of the simplest case (only IV x, but not DV y has measurement error), if the dependent variable y also has measurement error:
The observed coefficient can be either an overestimation or an underestimation of the true coefficient (Biemer&Trewin, p.619)

Methods for studying measurement errors (Gbook Ch7 → a great overview)
The above discussion is about how to conceptualize measurement error in a dataset. Now take a step back and think about 1) the general ways of capturing measurement errors and indicators of potential measurement errors and 2) how the above conceptualization and modeling link with survey design in practice.

A general comment is that researchers of bias often seek designs that shed light on cause of the error, and researchers of variance of seek designs that enable estimate of error and adjustment for inference (Lectur7)

1) **Laboratory experiments**
Do not specify a formal measurement error model;
Identify *causes* of measurement error
E.g., Cognitive interview (think-aloud, comprehension probes, debriefing interviews); Vignettes

Direct measures of cognitive steps
Questioning the respondents about his reactions to the survey questions:
"what does the word 'weekday mean to you?"
(p.342) This technique does not produce estimates of measurement error directly. Implicit in the work is the assumption that if two respondents mean different things by the words in the question, their answers to the question will have different amounts of error in them for measuring the intended concept.

Assumptions of laboratory investigation:
Heterogeneity of errors represented in the (small) subject pool
Measurement conditions are equivalent to those in the actual survey (attention level/motivation) External validity

2) **Measures external to the survey.**
Directly get to measurement errors of a survey by comparing survey results with external true values
a. Record check study—measurement error of individual responses

Reverse record check study
A sample is drawn from record file and interviews are taken containing questions about information contained on the record.
Example: Based on record, a population of rapists, draw a sample of rapist and interview their criminal activities.
Good for identify underreport (is a rapist but claim not) but cannot capture over-report (is not a rapist but claim yes),

because the people who do not have that traits are not included in the sample in the first place

Implicit measurement error model:

$$y_j = R_j + \varepsilon_j$$

where R_j is the record value for the respondent and is assumed to be the true value free of error (this assumption might not hold)

Measurement error is the difference between $E(R)$ and $E(y)$

Forward record check study

After survey responses are obtained in a sample, relevant record systems containing information on respondents are searched.

Example, ask a general population sample whether they are rapist. For the ones who report yes, search their record.

Good for identify overreport (claim yes, but is not a rapist), but cannot capture underreport because this method only searches the records for the people who reported yes.

Full design record check study

Combine the above two designs. Have a general sample and check records for all members of that sample (this essentially requires having records of the entire population).

Good for identify both underreport and overreport.

Common usage of record check—Good for identifying measurement *bias*. Using the records as the gold standard, record check studies can compare different measurement procedures and identify the one that produces results closest to the record (least biased results)

b. Comparing to population statistics—measurement error of survey estimates

Individual-level records are hard to obtained. As an alternative, validate survey statistics by comparing survey-based estimates of population parameters to the same parameters based on another (reliable) methodology.

But this is an umbrella indicator because the difference between the survey estimates and external population statistics is not only a result of measurement error, but also of coverage error and non-responses.

3) Randomized Assignment of measurement procedures to sample persons
Split sample experiment

—When *only two or least a small number of* alternative design features are compared
→ Can study biases

Under this framework, it makes sense to see measurement error as arising from, for example, *survey questions*—a *fixed* property of the design.

[If perform regression, the design features can be modeled as fixed effect]

An example:

"Do you think the US should forbid public speeches in favor of communism"

"Do you think the US should allow public speeches in favor of communism"

Implicit measurement model:

Method 1: $y_{1j} = \mu_{1j} + M_1 + e_{1j}$; Method 2: $y_{2j} = \mu_{2j} + M_2 + e_{2j}$

Difference of biases of the two methods:

$$M_1 - M_2 = Mean(\mu_{1j} + M_1 + e_{1j}) - Mean(\mu_{2j} + M_2 + e_{2j}) = \bar{y}_1 - \bar{y}_2$$

Interpenetration

—When *many* design features are compared

→ Can study variable errors

Under this framework, it makes sense to see measurement errors as arising from, for example, *interviewers*, a *variable* property of the design.

[If perform regression, the design features can be modeled as random effect → multilevel analysis]

Link to the HHB and ANOVA model discussed above:

$$y_{ji} = \mu_{ji} + b_i + e_{ji}$$

Intra-class correlation (i.e., intra-interviewer correlation):

$$\rho = \frac{\sigma_b^2}{\sigma_\mu^2 + \sigma_b^2 + \sigma_e^2}$$

This ***ICC*** inflates the variance of survey estimator

$$V(\bar{y}_{with\ design\ effect}) = V(\bar{y}_{assuming\ SRS})(1 + \rho(m-1))$$

where m is the averaged workload per interviewer

Note that if each interviewer work with only one case, then m equal to 1 and there is no design effect. Does it mean that by having each interviewer work on one case, we eliminate interviewer effect? No! In this case, b_i becomes b_{ij}, and it is no longer possible to differentiate b_{ij} and e_{ji}. The interviewer error merges into the simple response term rather than disappears.

Problems

There are problems with the randomization procedure:

a. Is it really possible that other errors are constant over assignment groups?
   For example, for a split ballot design, if the non-response rate of the two assignment groups differ, how does this affect the results?
   If the nonresponse rate of different interviewers differ, how does this affect the results?
b. "Hawthorne-like" effects
   External validity: Can behavior in each experimental group be replicated outside the experimental setting?
c. What if the magnitude of the measurement error associated with some unit in the design is dependent of the number of units used in the design?
   For example, the number of interviewers used can be negatively associated with interviewer quality (bc training cost). But then the more interviewers used on a project, the more stable are the estimates of interviewer variance. → paradox

4) Repeated measurement of the same person

The above randomization is about assigning different features to different people and assuming that by randomization, the different groups are equivalent. Here, it is about giving repeated measurements on the same person and investigate variability in response behavior over trails or replication. There are different definitions of

"repeated": can either achieve through test-retest or giving slightly different versions of the same measure

Interview-reinterview

A key premise is the stability of true values over trails (often within 1-4 weeks)

Response at trial 1 = true value + random error 1

$$y_{j1} = \mu_j + \varepsilon_{j1}$$

Response at trial 2 = true value + random error 2

$$y_{j2} = \mu_j + \varepsilon_{j2}$$

$E(\varepsilon_{j1}) = E(\varepsilon_{j2}) = 0$

$Cov(\varepsilon_{j1}, \varepsilon_{j2}) = 0$

A measure of reliability in response is provided by the "*index of inconsistency*" defined as the ratio of the variance of response errors to the total variance of the measure:

$$I = \frac{E_n(\varepsilon_{j1} - \varepsilon_{j2})^2/n}{\sigma_y^2} = \frac{E_n(y_{j1} - y_{j2})^2/n}{\sigma_y^2}$$

where σ_y^2 is variance over persons (j) of the observed values on y

In psychometric literature, *reliability* is assed in test-retest design by correlation of the two observed score:

$$\rho_{y_{j1},y_{j2}} = \frac{Cov(y_{j1}, y_{j2})}{\sqrt{Var(y_{j1})Var(y_{j2})}}$$

This is the reverse of I:

$$\rho_{y_{j1},y_{j2}} = 1 - I$$

For categorical variable, on the other hand:

		Trial 1	
		0	1
Trial 2	0	A	B
	1	C	D

Response variance can be indicated by gross difference rate $\frac{B+C}{N}$

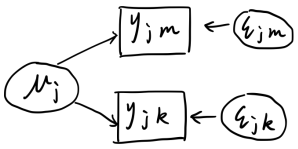
Problems:

Assuming no change in the true value over trials, threatened by long time between trials

Assuming no memory effect of the first trial on the second trial, threatened by short time between trials

Multiple indicators of the same construct

a. *Parallel measures*



$y_{jm} = \mu_j + \varepsilon_{jm}$

$y_{jk} = \mu_j + \varepsilon_{jk}$

μ_j is the true value of individual j

y_{jm} is response on item m

y_{jk} is response on item k

Assumptions of parallel measure are that:

1) $E(\varepsilon_{jm}) = E(\varepsilon_{jk}) = 0$ the two indicators have the same expected value, which is the true score
2) $Var(\varepsilon_{jm}) = Var(\varepsilon_{jk})$ the two indicators measure the construct equally well
3) Error terms are independent of the true value $Cov(\mu_j, \varepsilon_{jm}) = 0$

Conditioning on the two being parallel measures and errors independent, the reliability of y_{jm} and y_{jk} is:

$$\rho_{\mathrm{y}_{jm},\mathrm{y}_{jk}} = \frac{\mathrm{Cov}(\mathrm{y}_{jm},\ \mathrm{y}_{jk})}{\sqrt{Var(\mathrm{y}_{jm})\mathrm{Var}(\mathrm{y}_{jk})}}$$

Recall from above that given one item under Model 0, say $y_j = \mu_j + \epsilon_j$, reliability is defined as $\frac{\sigma_\mu^2}{\sigma_\mu^2+\sigma_\varepsilon^2} = \frac{\sigma_\mu^2}{\sigma_y^2}$. Given this definition, how to calculate the reliability? The parallel measure is one way to calculate this, see:

$$\frac{Cov(y_{jm}, y_{jk})}{\sqrt{Var(y_{jm})\,Var(y_{jk})}}$$

$$\begin{aligned}
Cov(y_{jm}, y_{jk}) &= E(y_{jm} \cdot y_{jk}) - E(y_{jm}) \cdot E(y_{jk}) \\
&= E[(\mu_j+\varepsilon_{jm})(\mu_j+\varepsilon_{jk})] - E(\mu_j+\varepsilon_{jm}) \cdot E(\mu_j+\varepsilon_{jk}) \\
&= E(\mu_j^2) + E(\mu_j \cdot \varepsilon_{jk}) + E(\mu_j \cdot \varepsilon_{jm}) + E(\varepsilon_{jm} \cdot \varepsilon_{jk}) \\
&\quad - E(\mu_j)^2 - E(\mu_j) \cdot E(\varepsilon_{jk}) - E(\mu_j) \cdot E(\varepsilon_{jm}) - E(\varepsilon_{jm}) \cdot E(\varepsilon_{jk}) \\
&\qquad\qquad 0 \qquad\qquad\quad 0 \qquad\qquad\quad 0 \\
&= E(\mu_j^2) - E(\mu_j)^2 \quad \leftarrow \text{bc } \mu_j \perp \varepsilon_{jm}/\varepsilon_{jk} \text{ and } \varepsilon_{jm} \perp \varepsilon_{jk} \text{ in model 0} \\
&= Var(\mu_j)
\end{aligned}$$

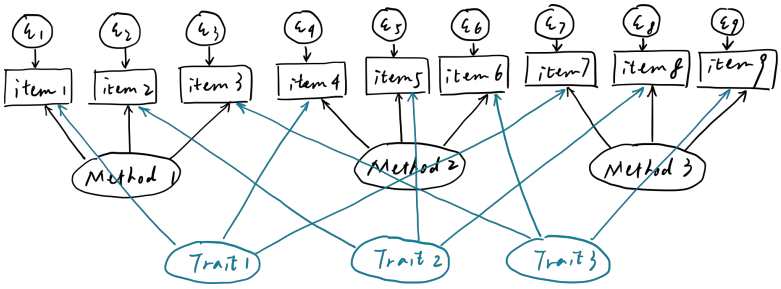
$\sqrt{Var(y_{jm}) \cdot Var(y_{jk})} = Var(y_{jm}) = Var(y_{jk})$ ← bc m & k are parallel measures

$$\therefore \frac{Cov(y_{jm}, y_{jk})}{\sqrt{Var(y_{jm})\,Var(y_{jk})}} = \frac{\sigma_\mu^2}{\sigma_y^2}$$

b. An extension to this is *multi-trait-multi-method (MTMM) model* (also in Saris)
   Partition variance in responses (across respondents) into variances in:
   - True values
   - Method effects
   - Simple random variances

Table 7.8 Nine Measurement Models for a Multitrait Multimethod Analysis with Three Methods and Three Traits, Estimated by Nine Indicators

Method 1	Trait 1	Item 1	$y_{i111} = \mu_1 + \beta_{11}X_{i1} + \alpha_{11}M_{i1} + \varepsilon_{i1}$
Method 1	Trait 2	Item 2	$y_{i122} = \mu_2 + \beta_{22}X_{i2} + \alpha_{12}M_{i1} + \varepsilon_{i2}$
Method 1	Trait 3	Item 3	$y_{i133} = \mu_3 + \beta_{33}X_{i3} + \alpha_{13}M_{i1} + \varepsilon_{i3}$
Method 2	Trait 1	Item 4	$y_{i214} = \mu_1 + \beta_{14}X_{i1} + \alpha_{24}M_{i2} + \varepsilon_{i4}$
Method 2	Trait 2	Item 5	$y_{i225} = \mu_2 + \beta_{25}X_{i2} + \alpha_{25}M_{i2} + \varepsilon_{i5}$
Method 2	Trait 3	Item 6	$y_{i236} = \mu_3 + \beta_{36}X_{i3} + \alpha_{26}M_{i2} + \varepsilon_{i6}$
Method 3	Trait 1	Item 7	$y_{i317} = \mu_1 + \beta_{17}X_{i1} + \alpha_{37}M_{i3} + \varepsilon_{i7}$
Method 3	Trait 2	Item 8	$y_{i328} = \mu_2 + \beta_{28}X_{i2} + \alpha_{38}M_{i3} + \varepsilon_{i8}$
Method 3	Trait 3	Item 9	$y_{i339} = \mu_3 + \beta_{39}X_{i3} + \alpha_{39}M_{i3} + \varepsilon_{i9}$



5) Collection of correlates of measurement error (e.g., paradata) (Lecture7)
Paradata are data collected as a byproduct of the survey process:

Measuring characteristics of the interview itself (Gbook Ch7)
Besides building models, observations of the interview situations either on site or through tape record can indicate the potential of measurement errors:

Did the interviewer follow the instruction?
Did the respondent ask for clarification?
Did the respondent hesitate before giving a response? Etc.

Do not posit a direct link between the behaviors observed and the measurement errors in the data. Rather, investigations appear to use the departures from specified procedures as prima facie evidence of measurement error.
For computer assisted surveys, can easily collect response behaviors, e.g., response latency, changes of responses etc.

Empirical studies examples
The above discussions are mostly theoretical, here are some empirical studies demonstrating the applications of these theoretical ideas.

Kreuter ~ Social desirability bias in CATI, IVR, and Web survey
2008 Investigates how different modes differ in their effect on socially desirable reporting (i.e., bias). Based on alumni survey, so there were records for all respondents to perform the full design record check to identify both over-report and under-report. Their dependent variables are categorical (e.g., at least one D/F, dropping a class etc.), so they calculate false positive rate and false negative rate to capture the biases.

Findings. They found 1) for reporting sensitive information: Web > IVR > CATI; the mode effect is larger for undesirable characteristics than for desirable ones. 2) They validated the assumption that an increase in reported undesirable behaviors points to higher accuracy. 3) They also pointed out whether questions about social undesirable

behaviors are indeed regarded as sensitive dependents on the respondents. Only the respondents who are in socially undesirable categories regard these questions as sensitive. This points to the possibility of correlated measurement error (ϵ_j) and true value (μ_j) (i.e., violation of the deduction of assumption 2 of Model 0).

What's special about this study? Typically, socially desirable responses are indirectly indicated by comparing different designs and viewing the ones with a higher report of undesirable behaviors and a lower report of desirable behaviors as closer to the truth. But this is based on an assumption. This study actually had record as the real truth and compare the effects of different modes. They also validate this widely adopted assumption.

Antoun ~ Effects of mobile versus PC web on survey response quality: A crossover experiment in a probability web panel

2017 Conducted a randomized crossover experiments to compare the effect of mobile web and PC web on the response quality using LISS panel. In wave 1, respondents were assigned to PC or mobile; in wave 2, the respondents were assigned to the opposite mode. Indicator of response quality (i.e., indicator of potential in measurement error) are non-differentiated answers, rounded numerical responses, short answers to open-ended answer, low cognitive reflection score, input errors (using a slider or date picker) + socially desirable response. Hypothesis is that mobile phone may result in lower quality data because people are more likely to be distracted and have low privacy when using phone.

Findings. Contrary to expectation, mobile did NOT elicit more satisficing responses than PC (equal conscientious and disclosure), even though people were more likely to be distracted (e.g., walking, and have people around) in mobile mode. But mobile did incur more input error when manipulating the slider and day picker → the small screen of the smartphone does make the interaction moderately difficult for certain kind of questions.

Interviewer effect

Interviewers are important for the measurement. Here are discussions dedicated to interviewer effect; some of these discussion overlaps with the above (Lecture9).

First, something confusing about modeling the interviewer effect. In the above section, based on Biemer&Strokes, HHB model in contrast to ANOVA model has be discussed. But the lecture notes present a different way of HHB model and contrast it to Kish model. And what exactly is HHB model is not presented in the same way… Anyway… This is what the lecture says:

ANOVA = Kish

HHB model: $y_{ij} = \mu_{ij} + M_i + \varepsilon_{ij}$.

Intra-interviewer correlation: $\rho = \frac{Var(M_i)}{Var(M_i)+Var(\varepsilon_{ij})}$. Thus ρ is the ratio of between-interviewer variance to the *total error variance*.

Kish model: $y_{ij} = \mu_{ij} + M_i + \varepsilon_{ij}$

Intra-interviewer correlation: $\rho = \frac{Var(M_i)}{Var(\mu_{ij})+Var(M_i)+Var(\varepsilon_{ij})}$. Thus ρ is the ratio of between-interviewer variance to the *total variance.*
Kish's ρ is easier to calculate.
Design effect: intra-interviewer correlation decreases the effective sample size and inflates the variance of the estimator

$deff = 1 + \rho(m-1)$, where m is the averaged workload
Effective sample size = $\frac{n}{deff}$
Variance of the estimator = $Variance\ if\ SRS * deff$
Example,

	m=10	m=20	m=50
ρ = .004	1.04	1.08	1.20
ρ = .009	1.08	1.17	1.44
ρ = .031	1.28	1.59	2.52
ρ = .072	1.65	2.37	4.53

As workload m increases, even a small ρ can seriously shrink the effective sample size.

Estimating interviewer variance (West&Blom)
To single out interviewer effect, require interpretation design:

Full interpretation—random assignment of units to interviewers. This is hard to implement.
Quasi—Random assignment of units to the currently working interviewers (telephone survey where interviewers work with different shifts). Interviewer effect can only be estimated within each shifts. The respondents who answer phones in the morning are very different from the ones that answer in the evening. The differences between morning and evening respondents are genuine and cannot be attributed to their interviewers.
Partial—adjacent geographic areas are pooled and at least two interviewers are randomly assigned to units in each of these areas.

If interpretation is fully satisfied, can fit an empty multilevel model or ANOVA model to estimate between-interviewer variance.
However, if interpretation is not well satisfied, can leverage multilevel model to remedy by including for example individual-level characteristics (differences at the lower/individual level explain differences at the higher/interviewer level).

Practical issues with interviewer variance model:

- In the calculation of design effect, an averaged workload m is used for an overall ρ. But in reality, workload might substantively matter; ρ is likely change for interviewers who work different workloads.
- Interpenetration could change essential survey condition. For example, interviewers might be purposefully selected to do refusal conversion.
- Practical issues of interpenetration: costs, interviewers quit etc.
- Confound measurement error with non-response error. Interpenetration can at most guarantee that the cases assigned to the interviewers are the same. But different interviewers recruit different respondents. Even if interpenetration design, interviewers can end up working with very different cases.
    - Example, West&Olson: interviewer variance for one variable (age at marriage) was largely driven by difference between interviewer in measurement error, but interviewer variance for another variable (age at

divorce) was largely driven by difference between interviewers in non-response.

What determines intra-interviewer correlation? (West&Blom, GbookCh8 p.374, Lecture9)
[Note that I put down how these factors influence interviewer variance *and response quality*!]

- Mode of interview on interviewer effect:
  - CATI < FTF
    less departure from script in centralized phone survey.
    less interviewer presentation
    computerized in general helps

- Features of survey questions on interviewer effect:
  - Attitudinal/opinion/ambiguous/subjective > factual/objective
  - Open-ended > closed-ended
  - Emotional/sensitive > neutral/non-sensitive
  - Difficult/complex skip patterns > simple

  These questions introduce more opportunities for interviewers to intervene, assist, and priming effect more important

- Interviewer behavior
  - Probing and providing feedback: increase response quality
  - Conversational/flexible/personal interviewer: increase response quality
  - Build rapport: decrease response quality

  Less empirical studies about how these behaviors influence interviewer variance. Perhaps a source of interviewer variance. This is also where interviewer experiences and expectations can make an impact.

- Interviewer training
  - General survey experience: mixed results on response quality
  - Current survey experience: mixed results on response quality

  Again, less clear how training and experiences influence interviewer variance.

- Interviewer demographics:
  - Race/ethnicity: less potentially defensive responses to racially sensitive questions
  - Age: mixed results. Mostly null findings
  - Gender: mixed results. Mostly null findings.

  Social norm and priming

- Respondent demographics
  - Gender: mixed results
  - Education: mixed results
  - Age: older respondents, larger interviewer effect

- Matching demographics:
  - Increase response quality

  But interviewer variance?

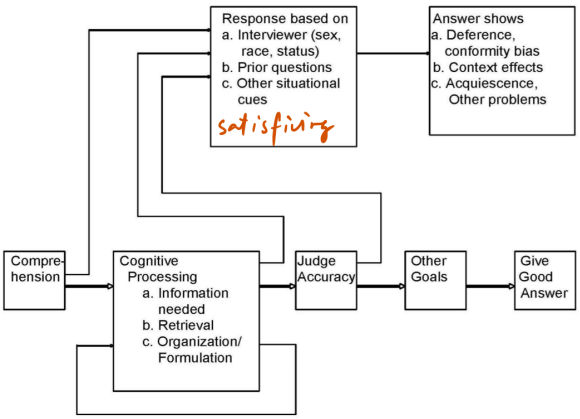
Now broaden the concept of interviewer effect. Rather than focusing on interviewer variance in specific, but look at interviewer effect in general. If interviewers influence response quality, what to do?

To lay the foundation, what should interviews be like and what role should interviewers play? The most prevailing philosophy and practice is standardized interviewing. Interviewers ideally are neutral tools and leave no traces in the data. The goal is to give the exact same stimuli to all respondents, so that the differences in data reflect genuine differences in respondents rather than the differences in the stimuli they receive.
To achieve this, interviewers always read the question exactly as it is worded and use only neutral probes.

In practice, however, interviewers deviate from guidelines (up to 14% with major changes); they may not sufficiently inform and motivate respondents; and they were sent to the field with incomplete training.

Three lines of research:

1) Cannel and colleague:
   High quality response is only possible when respondents stay motivated and committed. So if interviewers fail to play a good role of motivator, they have negative effects on the data quality.



   Solution is train interviewers to be good motivators: maintaining commitment, providing feedback, and providing instructions.
   Not much empirical support though…

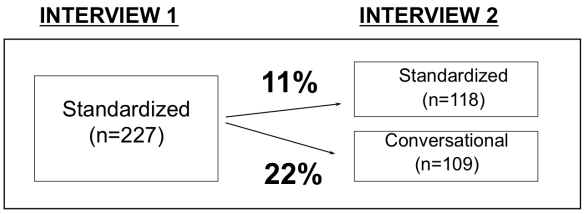
2) Fowler and Mangione:
   Proponent of strict standardized interviewing. Research on impact of training and supervision.
   No consistent evidence though. For example,
   Additional training make interviewers follow the guidance better, but no impact on ρ.
   Tighter supervision associated with lower ρ.

3) Conrad and Schober:
   Propose an alternative to standardized interviewing—conversational interviewing. Interviewers and respondents work together to assure respondent understands the question as intended. Even though this idea is presented as an alternative to standardized interviewing, it is virtually a variation of standardized: rather than standardizing wordings, conversational interviews aim to standardize meanings.

Example, two telephones interviews (longitudinal setting, 1 week in between) conducted on 227 respondents. In the first interview, all received standardized interview; in the second interview, half received standardized and half received conversational.

Finding.

- More responses changed when second interview was conversational than standardized

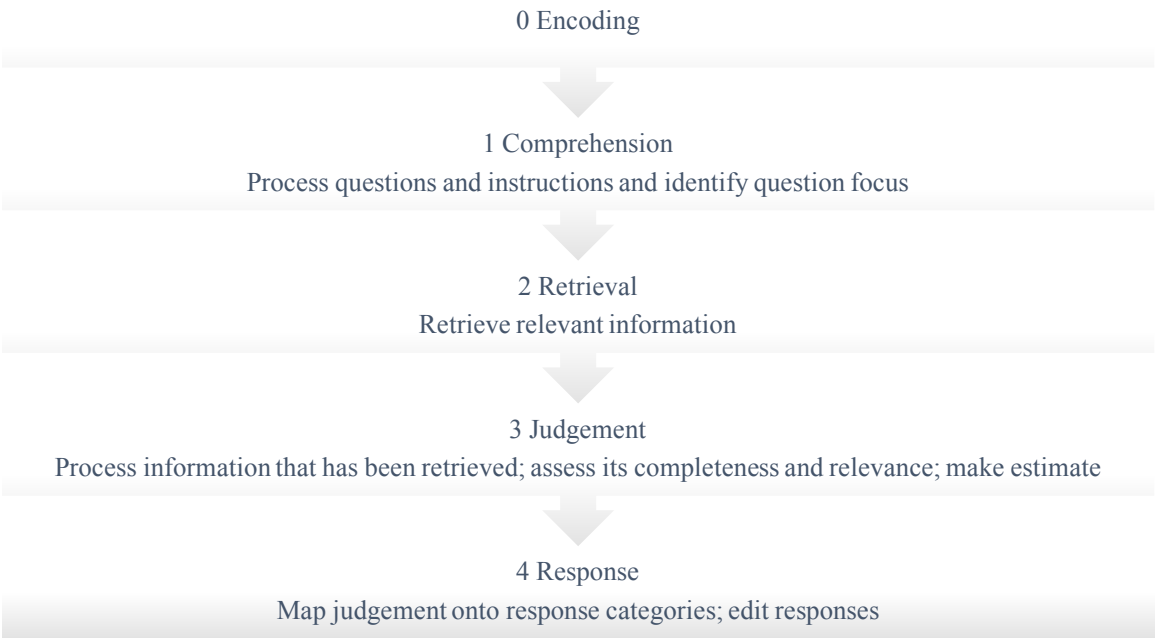


If conversation interviews help with respondents' comprehension, then responses should change more across the two waves than the ones consistently based on standardized interviews. Indeed the case.

Respondents and Questionnaires as the source of measurement error

This perspective dates back to the cognitive aspect of survey measurement movement in the 80s. It draws from cognitive and social psychology and looks into the causes that could generate response effect. The focus is how people respond to survey questions and what might result in errors, rather than consequences on survey measures. After this development, question testing becomes a standard practice.

First, in the **respondent side**, the most important things to know is the response process model (Lecture10)



- 0 Encoding
  This step is not included in the original response process model by Tourangeau. But later others argue that encoding—the process of storing information in long-term memory—should be the pre-requisite of all the following steps.

  One classical example demonstrating the importance of encoding is by Lee et al. 1999, on the problem of parents underreporting their child's vaccinations. The original hypothesis is that parents misreport because they forget. But the study demonstrated that parents' report immediately after the vaccination was as error-prone as their report 10 weeks later. This suggests that parents never encoded this information in the first place and there was nothing to be forgotten.

  Technically, encoding is more relevant to factual/event questions; but one line of research link encoding to attitudinal questions by viewing nonattitudes as lack of encoding. The problem here is that respondents mis-provide a substantive response even though they have no attitudes.
  One classical example is an experiment asking about people's attitudes using fictitious issues. What's your attitudes toward the "Agricultural Trade Act of 1978"? Respondents gave substantive responses even though this trade act is almost unheard of.
  Attempt to correct of this is to provide "DK" option or strength of attitudes questions.

- 1 Comprehension
  Comprehension involves analysis at various levels:
  - sensory level, lexical level (e.g., unfamiliar terms);
  - syntactic level (e.g., structural ambiguity);
  - semantic level (e.g., words have different meanings) ;
  - and pragmatic level (e.g., Conversations are guided by cooperative principle. People have expectations and fill in information depending on the context.)
  - Empirical examples:

    One specific example is maxim of relation:

    - Version 1: When riding in the back seat of a car, do you wear a seat belt all of the time, most of the time….
    - Version 2:
      - A. In the past year, have you ever ridden in the back seat of a car?
      - B. When you are riding in the back set of a car, do you wear a seat belt….

    10

Wear Seat Belt in Back Seat	One Question (%)	Two Questions (%)
All the time	30	42
Most of the time	27	16
Some of the time or once in a while	21	18
Never	24	4
Don't ride in back	8	20

    20% of people reported never wearing a seat belt even though they never sat in the back seat. This means that these respondents automatically modified the question to fit their own situation.

    Another specific example is assimilation and contrast effect—meaning of a sentence depends on what precedes it.
    General—"Happiness with life as a whole"
    Specific—"Happiness with marriage"

Context	Correlation	Explanation
General-specific	0.32	
Specific-general	0.67	Assimilation effect: bring in the information of the specific question to the

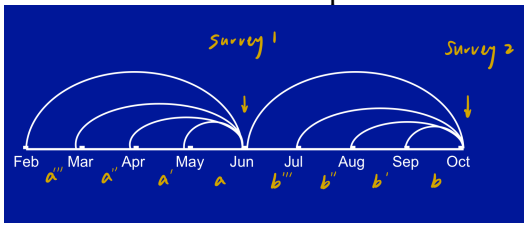
		judgement of the general question, thus high correlation
Joint lead-in	0.18	"Now we would like to learn about two areas of life that may be important for people's overall wellbeing." This explicitly contrasts the two aspects, thus low correlation

Considering framing examples

by Schuldt et al., 2011:
Republican—44% endorsed "global warming" and 60% endorsed "climate changes" as real
Democrats—unaffected by wording
This changes the conclusion. Using "global warming" would conclude that the partisan gap is large, but using "climate change" would conclude that partisan gap is small.

by Schuman and Presser:
"do you think the US should forbid public speeches against democracy"? → 39% yes forbid
"do you think the US should allow public speeches against democracy"? → 56% not allow

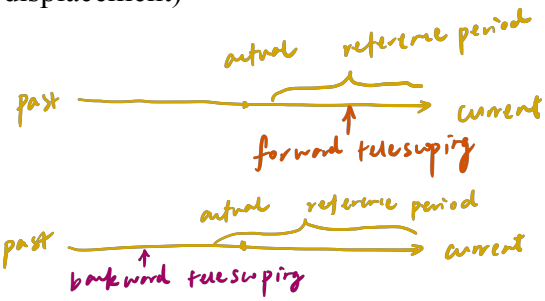
- 2 Retrieval
  - Retrieval strategies
    - Recall and count: tend to underestimate because of omit
    - Rate-based estimation: tend to overestimate because of overlooking exceptions
    - Impression-based estimation: tend to overestimate
  - Retrieval failure
    - Forgetting: memories decay most rapidly after the event experienced
    - Interference: the longer the time, the more likely that similar events occur, hard to distinguish details of one from another, memories all blend in together
  - Reconstruction error
    - Memories are reconstructive. The current knowledge influences memories in the past. It is almost impossible to accurately recall what one's original attitudes are. This is hindsight bias.
    - Seam effect. Specific to longitudinal surveys—respondents tend to underreport changes (up to 60%) between months covered by a single survey and over-report changes (up to 130%) between adjacent months included in the reference periods of different surveys



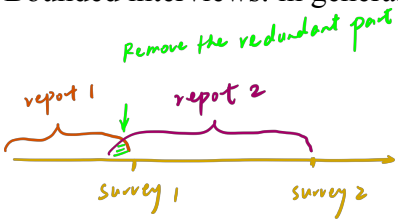
The result is sudden spikes in changes. This is a problem for longitudinal surveys that intends to track changes.

Why? Could be because of forgetting; because of interference (current knowledge interfering recall of the past); or because of "anchor and adjustment" (e.g., anchors at June and adjusts for May, April, March etc. tend to under-adjust)

- Telescoping—placing events in the wrong time frame (i.e., temporal displacement)



Methods to reduce telescoping;
Landmark events: public (Christmas) personal (birthday)
Calendars
Bounded interviews: in general, only work with panel interviews



Gaskell.etal. Study:
Landmark events are often used to help solve the telescoping problem. But are landmark events themselves subjected to telescoping? This study investigated telescoping of two landmark events—Thatcher resignation and football field disaster. There were indeed significant temporal displacements for both of the events: for forward telescoping Thatcher resignation and more backward telescoping for football disaster. No consistent pattern how demographic characteristics influence telescoping, except that women reported more forward telescoping.

- 3 Judgement
  Process the retrieved information and make estimate.
  Error arise when respondents use heuristics to form their estimates:
  - Availability.
    E.g., Assessment of risk influenced by salience of dramatic events
  - Anchoring and adjustment
    E.g., the seam example above is an example

  Opinion questions are error-prone because they often require on-site judgement. When respondents do not have preformed opinions, they have to either rely on a vague impression or construct evaluation.

- 4 Response
  Error occurs because of
  - social desirability—intentionally mis-report when the true response is embarrassing

- acquiescence response bias
- response categories do not well match with what the respondents have in mind. In addition, response categories carry information and might influence how the respondents report.
  For example:

  not at all successful |---|---|---|---|---|---|---|---|---|---| extremely successful
  0 10
  -5 +5

  When the scale ranges from -5 to 5, respondents do not like to choose the lower half of the scale because that implies failure.

Second, now the discussion moves on to **the questionnaire side**. The above discussion already touches a bit on this by showing that respondents react to the stimuli and different question wording/sequence might trigger different responses.

There are some conventional wisdoms about how to design questions in order to reduce error (Krosnick&Presser):

- Use 5- to 7- point scale
- Label all scale points (→ avoid scale ambiguity)
- Avoid agree/disagree questions (→ avoid ambiguity and acquiescence)
- Counterbalance order of response option (→ counterbalance order effect)
- Do not provide DK option (→ many respondents will use DK)
- Move from more general to more specific (→ else strong correlation between items, assimilation effect)
- Use multiple indicator of each construct (→ no single item is perfect, get to different dimensions of the construct)
- When measuring amounts (e.g., how many), use open question (→ different people interpret "often" differently)
- Ask contingent item only after all the screening questions have been administered (→ avoid speeding/satisficing. Respondents don't know that they can skip a lot of questions by answering "no")
- Reuse effective questions from earlier surveys

It is now a standard practice to perform question testing before largely fielding survey. There are different types of question-testing methods (Presser.etal & Maitland&Presser):

Some methods require data collection of the actual survey (smaller or larger scale). E.g.,

- Behavior coding—do interviewers have problems implementing certain questions?
- Response latency—do respondent hesitate when responding to certain questions?
- Statistical model—item response theory gets to how different items capture different levels of the latent construct

Some methods do not require such data collection. E.g.,

- Cognitive interviewers
- Expert review
- Computer-based evaluation: QUAID (question understanding AID); SQP (survey quality predictor); QAS (questionnaire appraisal system)

Discussing bias and variance (Applied sampling lecture)

Y_i denotes true characteristics of person i

y_{ij} denotes the observed value for the jth trial for person i

$\bar{y}_j = \frac{1}{n}\sum^{n} y_{ij}$ denotes the mean estimate of one trait.

$$MSE(\bar{y}_j) = E_j(\bar{y}_j - \bar{Y}_{true})^2 = E_j[(\bar{y}_j - \bar{E}_j(\bar{y}_j)) + (\bar{E}_j(\bar{y}_j) - \bar{Y}_{true})]^2$$

$$= \bar{E}_j[\bar{y}_j - \bar{E}_j(\bar{y}_j)]^2 + \bar{E}[\bar{E}_j(\bar{y}_j) - \bar{Y}_{true}]^2$$

$$= \underbrace{E_j[\bar{y}_j - \bar{E}_j(\bar{y}_j)]^2}_{Variance(\bar{y}_j)} + \underbrace{[E_j(\bar{y}_j) - \bar{Y}_{true}]^2}_{bias^2}$$

let $y_{ij} = Y_i + d_i + \varepsilon_{ij}$ ← random error differ from trial to trial

Y_i: true value of person i; d_i: systematic deviation of person i

$\bar{E}(y_{ij}) = Y_i + d_i$ is the consistent response one can expect from person i

$= C_i$ let me use C_i to denote it.

The overall setup should be considered as a two-stage sampling. We first sample individual (y_i) from the population → one layer of randomness; then the respondent select a random response ($+\varepsilon_{ij}$) → another layer of randomness

$$E(\bar{y}_j - \bar{E}_j(\bar{y}_j))^2 = \bar{E}(\bar{y}_j - C_j + C_j - \bar{E}_j(\bar{y}_j))^2 \quad \text{where } C_j = \frac{1}{n}\sum^{n} C_i$$

the mean of the consistent responses of the current sample

$$= \underbrace{E(\bar{y}_j - C_j)^2}_{E\left(\frac{1}{n}\sum^{n} y_{ij} - \frac{1}{n}\sum^{n} C_i\right)^2} + \underbrace{E(C_j - \bar{E}_j(\bar{y}_j))^2}_{\text{sampling variance variation of the mean of the consistent responses of the current sample from the mean across repetitions}} + 2E(y_j - C_j)(C_j - \bar{E}_j(\bar{y}_j))$$

$$E\left(\frac{1}{n}\sum^{n} e_{ij}\right)^2 = E(\bar{e}_j)^2 = V(\bar{e}_j)$$

$$= \sigma_c^2$$

This is why $C_j \neq \bar{E}_j(\bar{y}_j)$

C_j is the most we can get from the current sample of n respondents.

$\bar{E}_j(\bar{y}_j)$ is what we can get through repeated sampling people.

Alternatively:

$$E\left(\frac{1}{n}\sum^{n} e_{ij}\right)^2 = \bar{E}\left(\frac{(\sum^{n} e_{ij})^2}{n^2}\right) \quad \text{(be no covariance)} \quad = \bar{E}\left[\frac{1}{n}\frac{\sum^{n}(e_{ij}-0)^2}{n}\right]$$

$$= E\left[\frac{\hat{\sigma}^2_{\varepsilon_{ij}}}{n}\right] = \frac{\sigma^2_{\varepsilon_{ij}}}{n}$$

The above discussion is quite convoluted.

It can be expressed in more straightforward way:

$$y_{ij} = \underbrace{y_i + d_i}_{C_i} + \varepsilon_{ij}$$

where y_i is i's true value
and d_i is its systematic deviation
ε_{ij} is measurement error

given a sample of n respondents

$$V(\bar{y}) = V\left(\frac{1}{n}\sum^{n} y_{ij}\right) = \frac{1}{n^2} n \cdot V(y_{ij}) = \frac{1}{n} V(C_i + \varepsilon_{ij})$$

$$\therefore = \frac{\sigma_{C_i}^2}{n} + \frac{\sigma_{\varepsilon_{ij}}^2}{n}$$

let C_i here be a r.v. whose randomness comes from the sampling mechanism

$C_i = I_i C_i$

Total Survey Error
W11—Processing issues in surveys

Reference
Granquist and Kovar. 1997. Editing of survey data: How much is enough. → Granquist&Kovar
Campanelli et al., 1997. The quality of occupational coding in the United Kindom. → Campanelli.etal
Winkler. 1995. Matching and record linkage. → Winkler
Reiter. 2012. Research synthesis: statistical approaches to protecting confidentiality for microdata and their effects on the quality of statistical inference. → Reiter

Definition of data processing: "set of activities aimed at converting the survey data from its raw state as the output from data collection to a cleaned and corrected state that can be used in analysis, presentation and dissemination"

Data entry. Converting information recorded on a paper questionnaire to a computer-readable format: human keying, scanning, combination of the two.
Typical error rate of human keyers are very small (< 0.5%).

Editing
Survey data editing is the procedure for detecting (by means of edit rules) and for adjusting (manually or automatically) errors resulting from data collection or data capture.

The goal of editing should be three-fold: to provide information about the quality of the data, to provide the basis for the improvement of the survey vehicle, and to tidy up the data. However, a disproportionate amount of resources is concentrated on the third objective (Granquist&Kovar).

Types of edit checks (Lecture11):
- Range edits (age > 112 is false)
- Ratio edits (victim/crime < 1 is false)
- Comparison to historical data (given a two-month time gap, number of household members increase 5 is false)
- Balance edit (a+ b+ c < 100 is false)
- Consistency edits (age < 13, but married is false)

Real-time editing: With computer assistance, edit check can be performed during the data collection. E.g., CAI notifies the interviewer; web survey can program in hard edits or soft edits

Post survey editing (Granquist&Kovar)

Based on level of the aim, editing can be divided to
- *micro-editing* – ensuring validity and consistency of individual data records
- *macro-editing* – ensuring the reasonableness of data aggregates

Actions of the entire process include edit checks, re-contacting respondents (especially establishment surveys), alteration of entered data, and *imputation of missing data.

Based on certainty, editing can be classified to

fatal edits – identify data items that are certainly in error

query edits – flag data that look suspicious

Fatal edits should definitely be addressed. But too much attention is devoted to query. Given the large number of query edits flagged, hit rate is low and correction doesn't improve survey estimate in meaningful ways

Editing is costly:

- Monetary costs—manual reviewing by clerks and specialists; re-contacting respondents are expensive. In 75-80% of the survey, more than 20% of the total survey costs spent on editing.
- Increase respondent burden and introduce bad-will by querying respondents
- Lose timeliness
- Opportunity cost—could use the money to do other things
- Over-edits may lead analysts to rediscover the editor's model

Over-editing is really unnecessary. It can be counter-productive, introduce more error than remove.

Selective editing is one way out—not all errors are of equal importance, should give the errors that have actual impact priority and let go of the rest. Can cut down the current practice by 50%.

Coding

Survey coding can be seen as the process whereby textual information is classified into mutually exclusive categories and assigned numeric values (Campanelli.etal.)

Some terms in coding:

Responses—open-ended questions, half-open questions with "other" category

Nomenclature—code categories and code structures (e.g., occupation coding has levels of codes)

Coding instructions

Centralized coding vs. field coding

Manual coding vs. automated coding

Coding errors arise from:

1 Coding rules not properly applied

2 Ambiguity of correct coding number. E.g., what occupation to assign to survey professional?

3 Variation in judgements in coders

Coding errors can be reduced through verification:

Dependent verification: a second coder review the first coder's code, the second coder determines the final code

Independent verification: two coders code separately. In case of discrepancy, a third person resolves it

Campanelli.etal's study is one empirical example that looks at coding error. They report the quality of occupational coding (a complicated system—371 three digit codes) and investigated the effect of the computer-assisted options. In their study, reliability is defined as inter-coder consistency. Notice that consistency can be calculated both at the question level (agreement rate of two coders on this question) and the coder level (the proportion of codes that a person disagrees with others); validity is defined as the agreement of coders with experts. They differentiate between computer-assisted coding and computer-automated

coding.
They found only modest gain of computer-assisted coding comparing to manual coding; and that computer-automated coding is quite valid and reliable.
They also looked at coder variance (analogous to *interviewer variance*). Coder variance can be differentiated between
simple coder variance

$$y_c = \mu + \varepsilon_c$$

where coders introduce random error to the y_c. This cannot be identified in practice because ε_c will just merge with other simple random errors at the individual level (e.g., measurement error)

and *correlated coder variance*:

$$y_{cj} = \mu + b_c + \varepsilon_{cj}$$

where different coders introduce different systematic errors and thus there is correlation within each coder.
This can be addressed exactly like the interviewer variance. Interpenetration + having coded responses nested within coders → intra-coder correlation
Coder effect can again be calculated using the design effect equation:

$$codeff = 1 + \rho_c(m - 1)$$

Campanelli.etal found 1-1.13 codeff with workload of 300-400. Intra-coder correlation is small, but workload is large. Coding fewer questionnaires is better.

Disclosure avoidance
Trade-off between identification disclosure risk and data usefulness.

- For identification disclosure, many agencies based the risk measures on the probabilities that individuals can be identified from the released data. With microdata, Sweeney (2000) showed that 87% of U.S. population can be uniquely identified with only three pieces of information: 5-digit zip code, gender and date of birth. distance
- For data usefulness, it is usually assessed with statistical difference between the original and released data or differences of specific models between the original and the released data.

Methods of disclosure avoidance and their influence on survey analyses (Reiter):

- Recoding—collapse levels of categorical variables into fewer levels or cap the maximum. This is the most commonly used approach.
  Problem?—if analyses are at the same level as the collapsed data, then in general no problem. But if analyses intend at a different level, then might elicit ecological fallacy
- Data swapping—switching the data values for selected records with those for other records.
  Problem?—in general, the marginal distributions are retained, so if the goal is to estimate mean or total, swapping shouldn't be a big problem. But swapping may influence the relationships between swapped and un-swapped variables. However, if the swapped percentage is small (< 1%), the overall impact is minimal.
- Adding noise—add random noise to sensitive or identifying data values.
  Problem?—For continuous variables, point estimates remain unbiased although variance associated with those estimates may increase. For categorical variables, estimates are often biased. Theoretically, noise can be accounted for in analysis by

treating it similar as measurement error, but this requires knowing the noise distribution, which is not released by the agencies.

- Partially synthetic data—replace the original sensitive values with multiple imputations, multiple datasets are released to the public. (Develop models based on the raw data, then use the models to generate new datasets← leverage the superpopulation idea).
  Problem?—validity of synthetic data inferences depends on the validity of the models used to generate the synthetic data. Push the situation to an extreme: if the entire variables are simulated, analyses involving these variables can only reflect the models used to generate the data. Agencies often releases information about the synthesis models.

Matching and data linkage

This the opposite side of the story on disclosure avoidance. A large field of research look into how to link data of multiple sources, often between surveys and administrative data. Matching can be divided into exact matching based on unique identifiers, and statistical matching based on quasi-identifiers (assign matching weights to pairs and determine matching status by decision rules). Winkler study is one example showing the link of research on record linkage.

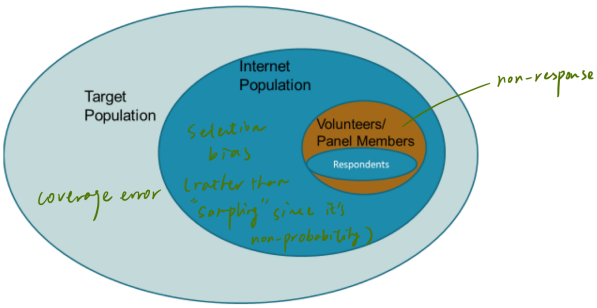
Total Survey Error
W12—Multiple sources of survey errors

Reference:
Tourangeau. 2018. How errors cumulate: Two examples. → Tourangeau lecture paper
Olson. 2006. Survey participation, nonresponse bias, measurement error bias and total bias. → Olson
Peytchev, Peytcheva, & Groves. Measurement error, unit nonresponse, and self-reports of abortion experiences. → Peytchev*Peytcheva
Sakshaug, Yan, & Tourangeau. 2010. Nonresponse error, measurement error and mode of data collection: Tradeoffs in a multi-mode survey of sensitive and non-sensitive items. → Sakshaug.etal

Tourangeau lecture paper:
The leading question is how different errors add up? Do they have different signs and thus cancel out each other, or do they have the same sign and thus the overall error is accentuated?

Consider two examples.

First, non-probability online web panel suffers from *coverage error*, *selection bias* and *non-response error*. How do the three errors add up?



$$B(\hat{\theta})_{tot} = B_{Int} + B_{sel} + B_{nr},$$
$$B_{cv} = \theta_{Int} - \theta_{full},$$
$$B_{sel} = \theta_{sam} - \theta_{Int},$$
$$B_{nr} = \theta_{res} - \theta_{sam},$$
$$B(\hat{\theta})_{tot} = (\theta_{Int} - \theta_{full}) + (\theta_{sam} - \theta_{Int}) + (\theta_{res} - \theta_{sam}).$$

- Coverage error ($B_{cv} = \theta_{Int} - \theta_{full}$) arises because not everyone has access to the internet. There are age, education, and race difference in internet access.
- Selection error ($B_{sel} = \theta_{sam} - \theta_{Int}$) arises because not everyone has access to the internet would opt for the web panel. In fact, these people seem to have very different characteristics. They do not value their time as much and they do a lot surveys.
- Non-response ($B_{nr} = \theta_{res} - \theta_{sam}$) arises because not everyone receives the invitation would complete the survey. Non-response is a vague term when it comes to non-probability web surveys. In the context of web intercept or banner recruitment, it is unclear who receive the invitation in the first place so it is impossible to calculate response rate. A more realistic context for calculating nonresponse rate is web panel. In general, web survey are prone to higher level of unit nonresponse than other surveys and a higher rate of breakoff.

How do these errors of non-probability web survey cumulate?

The speculation is that these error sources point to the same direction rather than cancel each other out. As a result, the overall bias as a function of the covariance between response propensity P and variable of interes Y (recall $Bias = \frac{Cov(P,Y)}{\bar{P}}$) becomes larger as these errors cumulate because now

$$Cov(p_{over}, Y) \geq Cov(p_{cov}, Y) \geq Cov(p_{pan}, Y) \geq Cov(p_{part}, Y).$$

Can weighting solve the problem?
Summarizing across studies—

- Weighting removes some of the bias, at most around $\frac{3}{5}$
- Weights sometimes make things worse → increase bias
- After weighing, some biases remain substantial, shifting the estimates by 20%
- The impact of weighting varies across estimates

Second, relationship between non-response error and measurement error. Would it be possible that decreases in non-response errors lead to increase in measurement errors?

The idea behind is that reluctant respondents (who would be non-respondents without extensive recruitment) might give sloppy responses (high measurement error) when are indeed successfully recruited. If this is the case, the effort in reducing non-response error might lead to increase in measurement error. Theoretical hypotheses explaining this link are:

- Reactance (multiple contact attempts provoke the respondents → worse responses)
- Topic and sponsorship (lack of interest → don't want to participate at first → worse responses once recruited)
- Research importance (similar to the above one, don't think research is important → don't want to participate at first → worse responses once recruited)
- Respondent characteristics (e.g., education → don't want to participate at first → worse responses once recruited)
- Self-perception (don't want to participate → don't perceive self as good a respondent → worse responses once recruited)
- Change in survey protocols (often need a change in the survey protocols to get the reluctant respondents → difference in response quality)
- Social desirability (socially undesirable groups → don't want to participate → misreport once participate)
- Commitment (having decided to participate increases motivation to do a good job – note: don't know why is this factor relevant…)

What is the empirical evidence with regard to this link between non-response error and measurement error?

There are many different indicators of measurement error. The general conclusion is that reluctant respondents have more item non-response, but they do not provide lower quality data with respect to other indicators.

Two other papers of this week specifically look into this link:

Empirical study on this issue I: Olson used Wisconsin Divorce Study—a study that sampled from divorce records and thus true values are available:

Sample records > respondents' records → difference → non-response bias
respondents' records > respondents' responses → difference → measurement bias

She classified respondents into 5 categories based on their response propensity and compared non-response bias and measurement bias of these 5 categories.

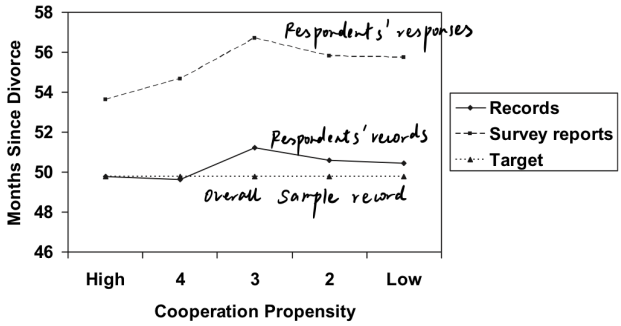
The goal is to address two research questions:

1) Would incorporating hard-to-get respondents influence MSE?

   Findings. How the means (based on records) change with and without including the hard to get respondents → there appear to be nonresponse error (the non-respondents have different records), but the overall non-response error is small.

2) Responses of the hard-to-get respondents are sometimes used to indicate non-response bias. Is it accurate to do so? What if the difference between hard and easy respondents actually reflects changes in measurement error?

   Now want to see how measurement error changes along with response propensity (difficult respondents have more measurement errors?).
   Here is one example of the result plots:



   As can be seen, there is measurement error (difference between dashed line and solid line), but the measurement error did not change across different cooperation propensities. So <u>there is little evidence that hard-to-get respondents provide inaccurate data.</u>
   If we want to know non-response bias, ideally, we look at the solid line (respondents' records) and see how low propensity respondents' records differ from the high propensity respondents' records. But of course, records are not available most of the time. The good news is that the dashed line *tracks* the solid lines closely. Thus, if the trend is what interests us, then looking at the dashed line is valid.

Empirical study on this issue II: Peytchev*Peytcheva investigated the link between non-response error and measurement error on the self-report of abortion.
Method-wise, they don't have the true abortion records (of course…). But the NSFG survey asked the abortion question both in ACASI and CAPI, so they assumed that ACASI's responses reflect the true values.

Finding. <u>Respondents, who were less likely to participate in the survey, were also more likely to underreport abortion.</u>

	Response Propensity Quintile				
	1 (Low)	2	3	4	5 (High)
Under-report	23.7%	13.2%	10.3%	13.3%	11.8%

Thus, this study support that recruiting hard-to-get respondents introduce more measurement error into the survey.

Empirical study on this issue III: Sakshaug.etal. used the Maryland alumni survey, which used three modes CATI, IVR and Web. This is again a study that had true values. They extended the investigation of the relationship between nonresponse error and measurement error by linking these two error sources to different modes. Now (one of) the questions have changed from how measurement error and nonresponse error trade off (decrease nonresponse error →? increase measurement error) to how different modes trade off (decrease measurement error but increase nonresponse error, worth it?). The research background is that self-administrated modes are good because they promote honest responses (low measurement error); but because of the way respondents are recruited, there is a switch in mode which results in a rather large percentage of drop out (high non-response rate).

The goal is to address three questions:

1) What was the relative contribution of nonresponse and measurement error to the overall error in the survey estimates?
   Findings. For questions on undesirable characteristics, there were rather large measurement biases. Measurement errors contribute more to survey estimates than non-response bias. For questions on desirable or neutral characteristics, measurement errors were no longer strikingly higher, and nonresponse errors actually became larger than the measurement errors.
2) Does the reduction in measurement error offset any increase in nonresponse bias due to the relatively large number of cases dropping out during the switch from an interviewer-administered mode to a self-administrated mode?
   Not really. The privacy of self-administrated mode did decrease measurement error, but the non-response error due to dropout increased by a larger extent. This increase in non-response error (sometimes, not always significant) canceled out the advantage of the self-administration.
3) How did the level of error needed to contact the sample members and get them to complete the screener related to the level of accuracy in their answers?
   No evidence that difficult respondents provide less accurate data.
   How did the difficult respondents' inclusion influence non-response bias? Small impact (meaning that these difficult respondents did not have very different records comparing to the easy respondents), but still a net gain.