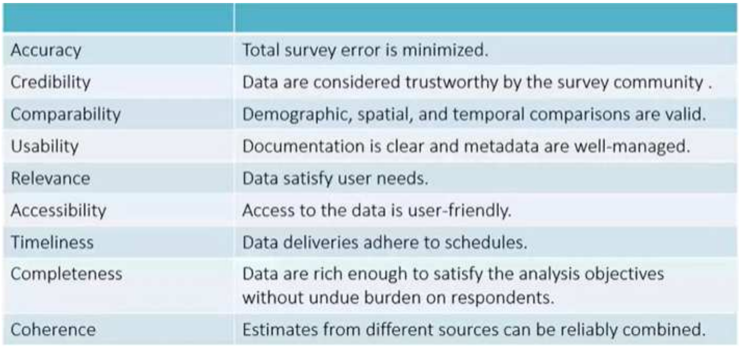
Framework for Data Collection and Analysis

Week 3 Quality Framework

Quality of Data

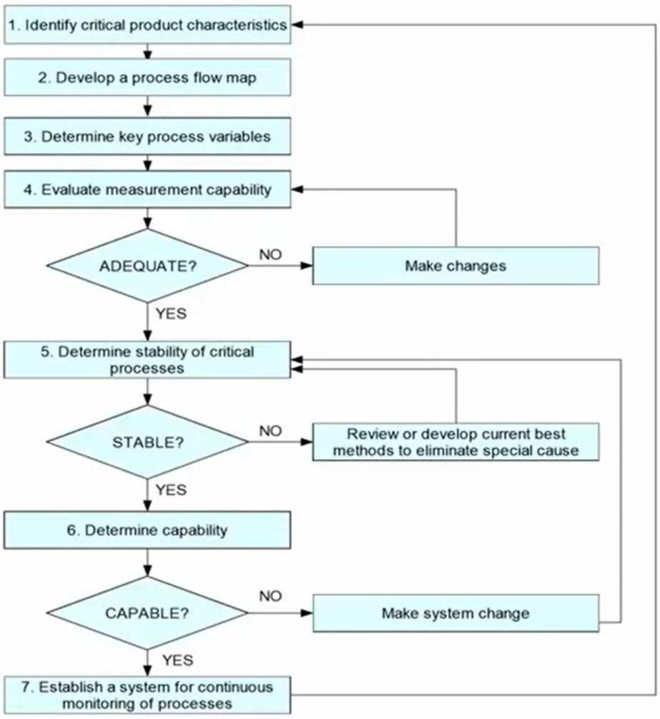
Overall Quality Assessment



Accuracy	Total survey error is minimized.
Credibility	Data are considered trustworthy by the survey community .
Comparability	Demographic, spatial, and temporal comparisons are valid.
Usability	Documentation is clear and metadata are well-managed.
Relevance	Data satisfy user needs.
Accessibility	Access to the data is user-friendly.
Timeliness	Data deliveries adhere to schedules.
Completeness	Data are rich enough to satisfy the analysis objectives without undue burden on respondents.
Coherence	Estimates from different sources can be reliably combined.

Continuous Quality Improvement

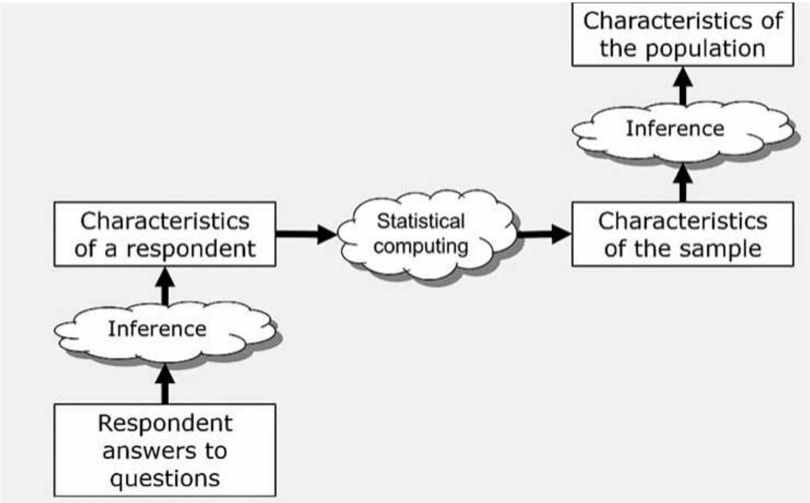
Prepare a workflow diagram of the process + identify key process variables
Identify characteristics of the process that are critical to quality
Develop real-time, reliable metrics for the cost and quality of each
Verify that the process is stable (i.e. in statistical control) and capable (i.e. can produce the desired results)
Continuously monitor costs and quality metrics during the process
Intervene as necessary to ensure that quality and costs are within acceptable limits
Morganstein and Marker (1997) Flow Chart for Quality Checks


1. Identify critical product characteristics
2. Develop a process flow map
3. Determine key process variables
4. Evaluate measurement capability
ADEQUATE?
NO
Make changes
YES
5. Determine stability of critical processes
STABLE?
NO
Review or develop current best methods to eliminate special cause
YES
6. Determine capability
CAPABLE?
NO
Make system change
YES
7. Establish a system for continuous monitoring of processes

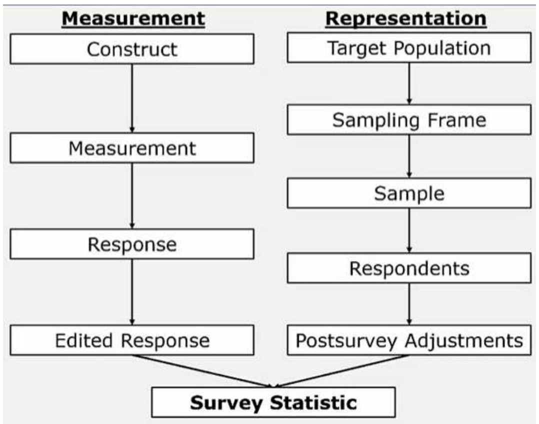
Inference

Two Types of Inference

Inference from response to a characteristic of a respondent (measurement)
Inference from the sample to the population (representation)


Characteristics of the population
Inference
Characteristics of a respondent
Statistical computing
Characteristics of the sample
Inference
Respondent answers to questions

Survey Lifecycle – *Design* Perspective



Measurement

- Construct: Elements of information sought by researcher, usually described by words, often abstract, permitting different more specific definitions

e.g. belief in God, happiness, quality of time, crime

- Measurement: Linking theoretic constructs to observable variables (ways to gather info about constructs; procedures operations step-by-step protocols implemented to gather data)

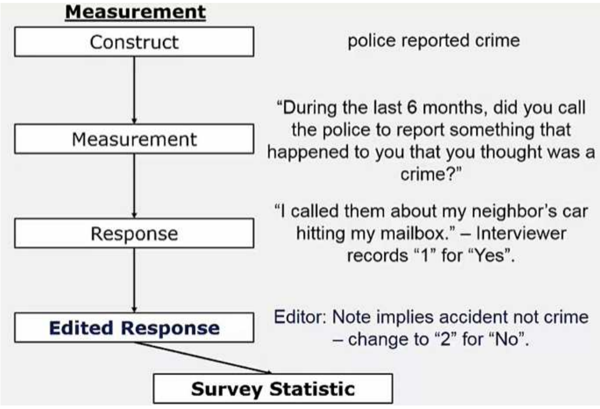
e.g. questions, observations, soil samples, blood pressure readings, blood drawings, observation of mouse movements on a website, survey questions

- Response: Respondent outcomes from measurements

e.g. answers to questions, quantity of soil, number from blood pressure device, record extraction

- Edited Response: Value stored in data record used for analysis for specific measure

e.g. resulting from coding (text to numbers), acceptable answer set (e.g. range edits), or consistency rules (e.g. contingency edits)



Measurement

Construct — police reported crime

Measurement — "During the last 6 months, did you call the police to report something that happened to you that you thought was a crime?"

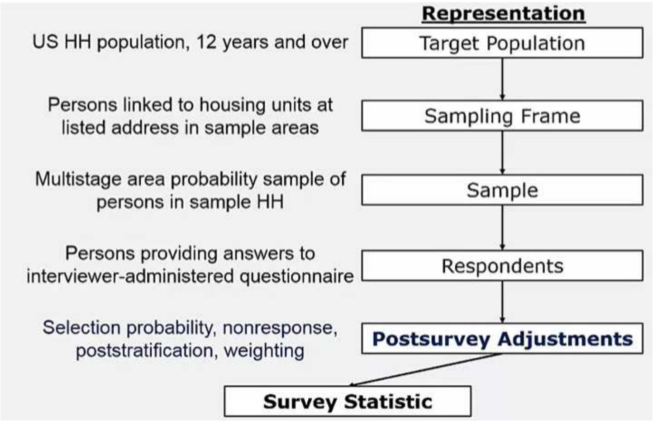
Response — "I called them about my neighbor's car hitting my mailbox." – Interviewer records "1" for "Yes".

Edited Response — Editor: Note implies accident not crime – change to "2" for "No".

Survey Statistic

Representation

- Target Population: Set of units to be studied; abstractly defined so that there are several ways to operationalize set
e.g. adults in the U.S. ; U.S. household population in 2015
- Sampling Frame: set of units identified in some way that it could be sampled and located; with an ideal sampling frame every unit in target population appears on sampling frame once and only once, and nothing else appears on sampling frame
e.g. telephone numbers as a frame for persons, list of dwellings created for area probability sampling
- sample: subset of frame population chosen for measurement in survey
- respondents: sample units that are successfully contacted and measured
- post survey adjustments: changes to records in survey data set to make survey estimates based on them better reflect full target population
e.g. selection weights, imputation, nonresponse weights, post stratification



Representation

US HH population, 12 years and over — Target Population

Persons linked to housing units at listed address in sample areas — Sampling Frame

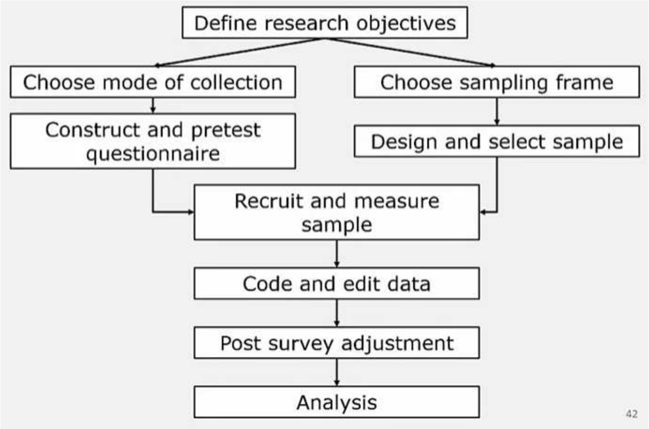
Multistage area probability sample of persons in sample HH — Sample

Persons providing answers to interviewer-administered questionnaire — Respondents

Selection probability, nonresponse, poststratification, weighting — **Postsurvey Adjustments**

Survey Statistic

Survey Lifecycle – *Process* Perspective



Define research objectives

Choose mode of collection

Construct and pretest questionnaire

Choose sampling frame

Design and select sample

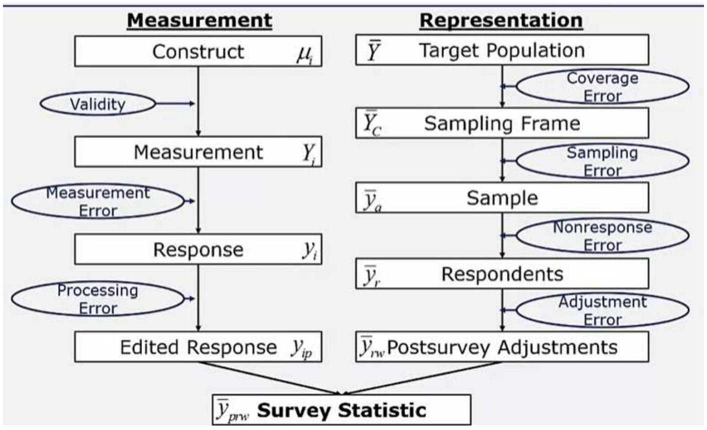
Recruit and measure sample

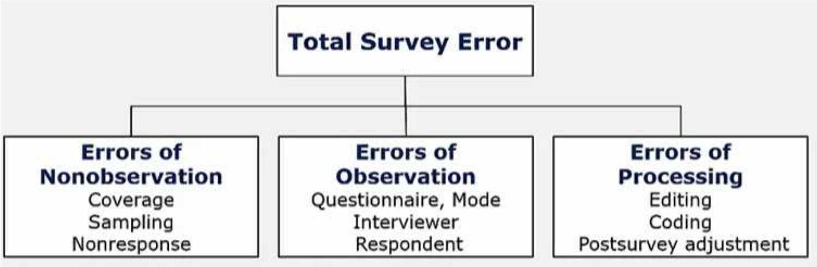
Code and edit data

Post survey adjustment

Analysis

Survey Lifecycle – *Quality* Perspective





Total Survey Error

· Concept, way of thinking about various sources of error that may affect survey statistics

· Error ≠ mistake but rather reflects uncertainty (or lack of confidence) of inference

· Survey quality / value

- minimize error for a given investment
- Success of TSE approach depends on good information on costs and errors

· Need to assess level of error associated with alternative procedures and choose combination of approaches best suited to problem

· Survey errors can arise from many sources

- survey topic, available funding, sampling frame, data collection method, interviewer training, etc.

· In sum, notion of TSE guides design decisions

- TSE framework helps understanding potential impact of design decisions on survey errors
- Together with costs, explicit part of design decisions

· Statistical notion of error is expressed as mean square error (MSE)

- MSE(y_bar) = (E(y bar) – Mu)^2
- Squared sum of all variable errors and biases
- Errors are specific to a certain statistic or estimate
- In practice, MSE rarely fully measured

· we need to have a Total Error perspective not only for designed but also for found data

True Value and Error

· True Value: An idealized concept of a quantity which is to be measured…

· Theoretic view: absolute standard for comparison even if not knowable

· Operational view

- True value defined in terms of measurement process
- e.g. IQ is outcome of a psychological assessment; predicts performance across vast range of tasks

Variable Error and Bias

· Variable Error

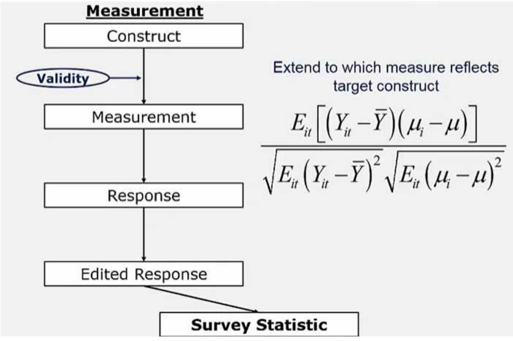
- Variation over replication

- Often represented by variance of a statistic
- Arises because achieved values vary over different units in the design (e.g. interviewers, sample persons, questions)
- Estimated from sample itself, using replication based methods

· Bias

- Systematic deviation from "true value"
- Directional Error
  - · e.g. bigger reports than are actually the case
- To estimate bias requires external data ("truth") or assumptions about direction of effects

Metrics



Measurement

Construct

Validity

Measurement

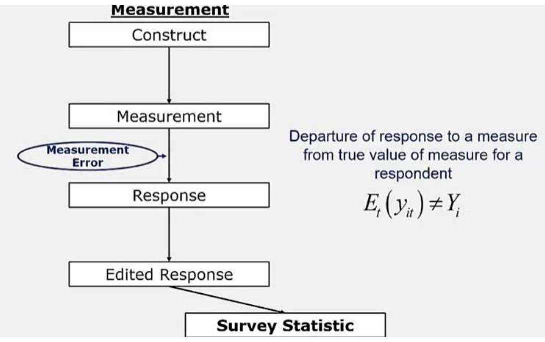
Response

Edited Response

Survey Statistic

Extend to which measure reflects target construct

$$\frac{E_{it}\left[\left(Y_{it}-\bar{Y}\right)\left(\mu_i-\mu\right)\right]}{\sqrt{E_{it}\left(Y_{it}-\bar{Y}\right)^2}\sqrt{E_{it}\left(\mu_i-\mu\right)^2}}$$



Measurement

Construct

Measurement

Measurement Error

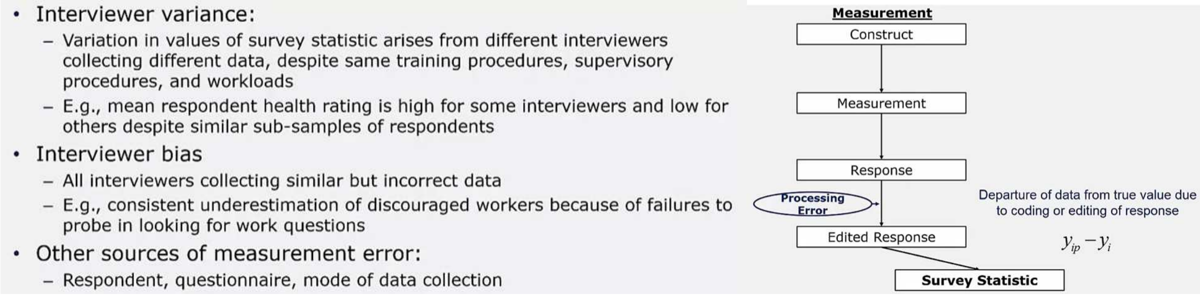
Response

Edited Response

Survey Statistic

Departure of response to a measure from true value of measure for a respondent

$$E_t\left(y_{it}\right) \neq Y_i$$



- Interviewer variance:
  - Variation in values of survey statistic arises from different interviewers collecting different data, despite same training procedures, supervisory procedures, and workloads
  - E.g., mean respondent health rating is high for some interviewers and low for others despite similar sub-samples of respondents
- Interviewer bias
  - All interviewers collecting similar but incorrect data
  - E.g., consistent underestimation of discouraged workers because of failures to probe in looking for work questions
- Other sources of measurement error:
  - Respondent, questionnaire, mode of data collection

Measurement

Construct

Measurement

Response

Processing Error

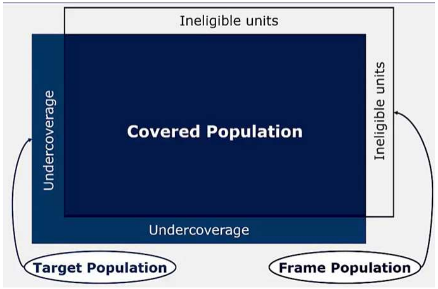
Edited Response

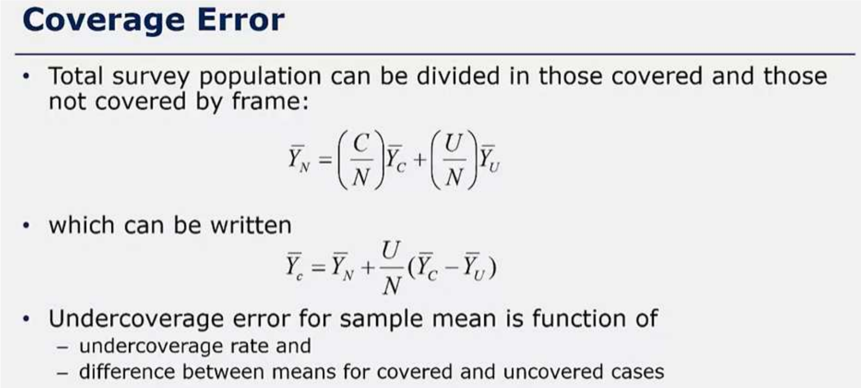
Survey Statistic

Departure of data from true value due to coding or editing of response

$$y_{ip}-y_i$$

Coverage and Sampling





Coverage Error

- Total survey population can be divided in those covered and those not covered by frame:

$$\bar{Y}_N = \left(\frac{C}{N}\right)\bar{Y}_C + \left(\frac{U}{N}\right)\bar{Y}_U$$

- which can be written

$$\bar{Y}_c = \bar{Y}_N + \frac{U}{N}(\bar{Y}_C - \bar{Y}_U)$$

- Undercoverage error for sample mean is function of
  - undercoverage rate and
  - difference between means for covered and uncovered cases

Coverage error is **(U/N)(Y_c bar 0 Y_u bar)** in the above equation

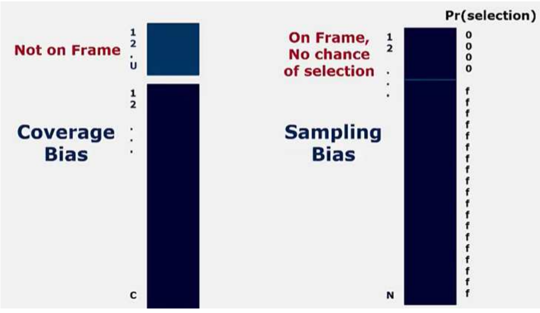
· Sampling variance

- Variation in values of survey statistics because different subsets of the population fall into sample over replications of same sample design
- Most commonly measured statistic in surveys
- Confidence intervals, standard errors

· Sampling bias

- Consistent failure to estimate a proportion of population
  e.g. those in military in HH samples which exclude military bases from every sample
- sampling bias is 0 for probability samples

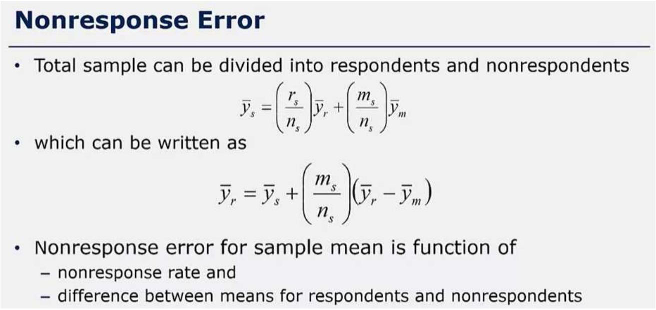
Coverage Bias (not covered in the first place) v.s. Sampling Bias (covered by selection probability = 0)



Non response Error and Adjustment Error

· Unit nonresponse: complete absence of an interview from a sampled household

· Item nonresponse: absence of answers to specific questions in the interview after the sampled household agrees to participate in the survey.



Nonresponse Error

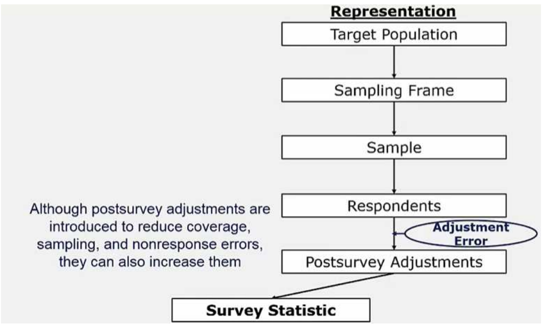
- Total sample can be divided into respondents and nonrespondents

$$\bar{y}_s = \left(\frac{r_s}{n_s}\right)\bar{y}_r + \left(\frac{m_s}{n_s}\right)\bar{y}_m$$

- which can be written as

$$\bar{y}_r = \bar{y}_s + \left(\frac{m_s}{n_s}\right)(\bar{y}_r - \bar{y}_m)$$

- Nonresponse error for sample mean is function of
  - nonresponse rate and
  - difference between means for respondents and nonrespondents



Representation

Target Population

Sampling Frame

Sample

Respondents

Adjustment Error

Postsurvey Adjustments

Survey Statistic

Although postsurvey adjustments are introduced to reduce coverage, sampling, and nonresponse errors, they can also increase them

· Variable Errors; Systematic Errors

- Some errors common to all trials of survey for given statistic (e.g. coverage bias due to missing cellphone-only HH in landline CATI survey)
- some errors vary over trials (e.g. variable response errors)

· There are no good or bad surveys – only good or bad survey statistics

- Errors are properties of statistics (e.g. sample mean is biased estimate of target population mean)
- From sample survey, some statistics may have large errors; others, mall errors

· Survey methodology research discovers how to reduce errors and applies these to surveys

Three areas of Survey Research by *Robert M. Groves*

· Survey Research three distinct era / stages

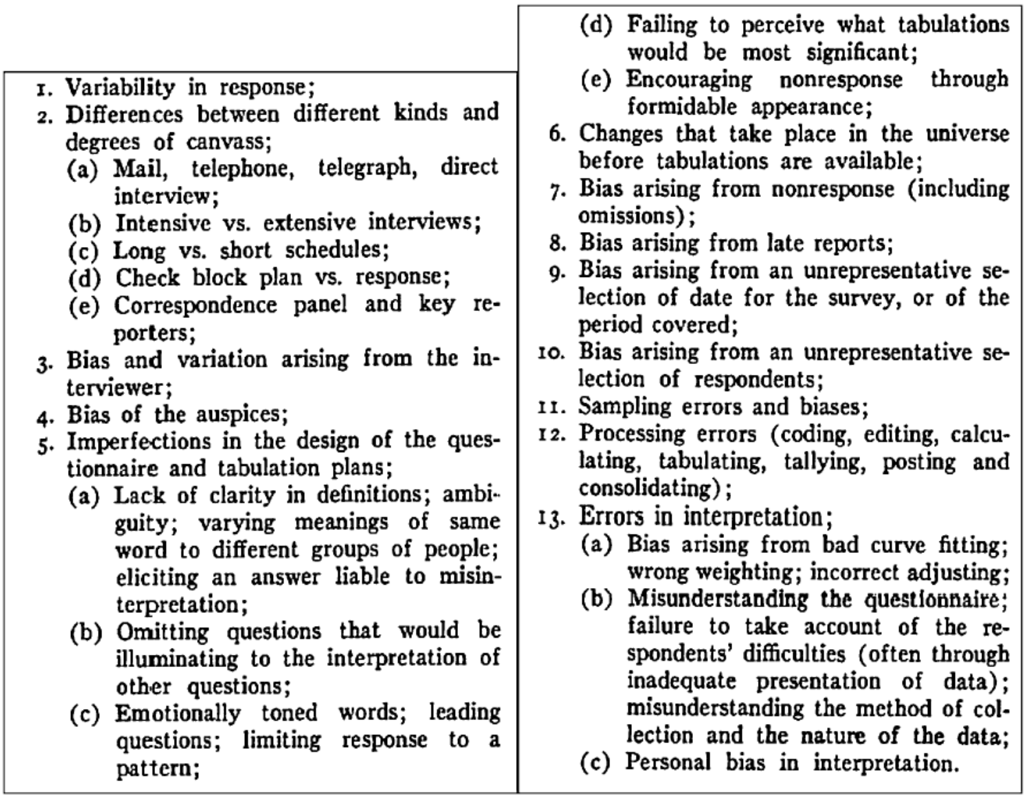
- First Era (1930–1960): the founders of the field invented the basic components of the design of data collection and the tools to produce the statistical information from surveys. As they were inventing the method, they were also building the institutions that conduct surveys in the private, academic, and government sectors.

Second Era

- Second Era (1960–1990): witnessed a vast growth in the use of the survey method. This growth was aided by the needs of the U.S. federal government to monitor the effects of investments in human and physical infrastructure, the growth of the quantitative social sciences, and the use of quantitative information to study consumer behaviors.

- The third era (1990 and forward): witnessed the declines in survey participation rates, the growth of alternative modes of data collection, the weakening of sampling frames, and the growth of continuously produced process data from digital systems in all sectors, but especially those emanating from the Internet.

[Article] TOTAL SURVEY ERROR PAST, PRESENT, AND FUTURE
by ROBERT M. GROVES LARS LYBERG



1. Variability in response;
2. Differences between different kinds and degrees of canvass;
   (a) Mail, telephone, telegraph, direct interview;
   (b) Intensive vs. extensive interviews;
   (c) Long vs. short schedules;
   (d) Check block plan vs. response;
   (e) Correspondence panel and key reporters;
3. Bias and variation arising from the interviewer;
4. Bias of the auspices;
5. Imperfections in the design of the questionnaire and tabulation plans;
   (a) Lack of clarity in definitions; ambiguity; varying meanings of same word to different groups of people; eliciting an answer liable to misinterpretation;
   (b) Omitting questions that would be illuminating to the interpretation of other questions;
   (c) Emotionally toned words; leading questions; limiting response to a pattern;
   (d) Failing to perceive what tabulations would be most significant;
   (e) Encouraging nonresponse through formidable appearance;
6. Changes that take place in the universe before tabulations are available;
7. Bias arising from nonresponse (including omissions);
8. Bias arising from late reports;
9. Bias arising from an unrepresentative selection of date for the survey, or of the period covered;
10. Bias arising from an unrepresentative selection of respondents;
11. Sampling errors and biases;
12. Processing errors (coding, editing, calculating, tabulating, tallying, posting and consolidating);
13. Errors in interpretation;
   (a) Bias arising from bad curve fitting; wrong weighting; incorrect adjusting;
   (b) Misunderstanding the questionnaire; failure to take account of the respondents' difficulties (often through inadequate presentation of data); misunderstanding the method of collection and the nature of the data;
   (c) Personal bias in interpretation.

Figure 1. Deming's Listing of 13 Factors That Affect the Usefulness of a Survey (1944).

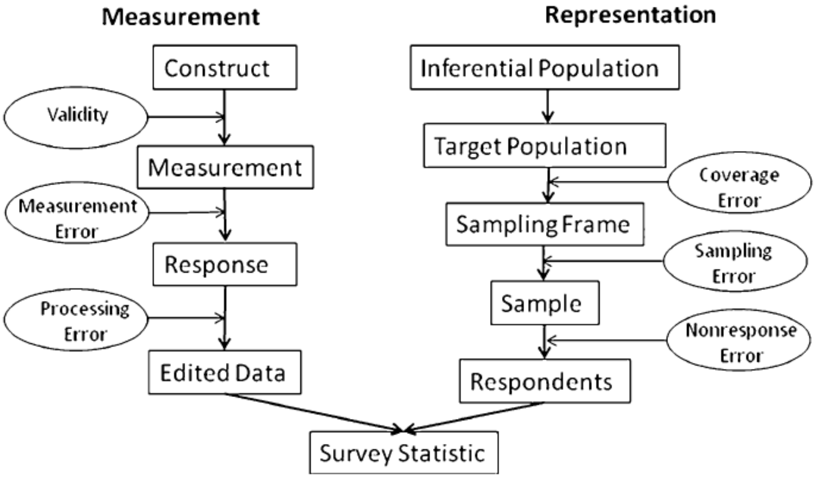


Figure 3. Total Survey Error Components Linked to Steps in the Measurement and Representational Inference Process (Groves et al. 2004).

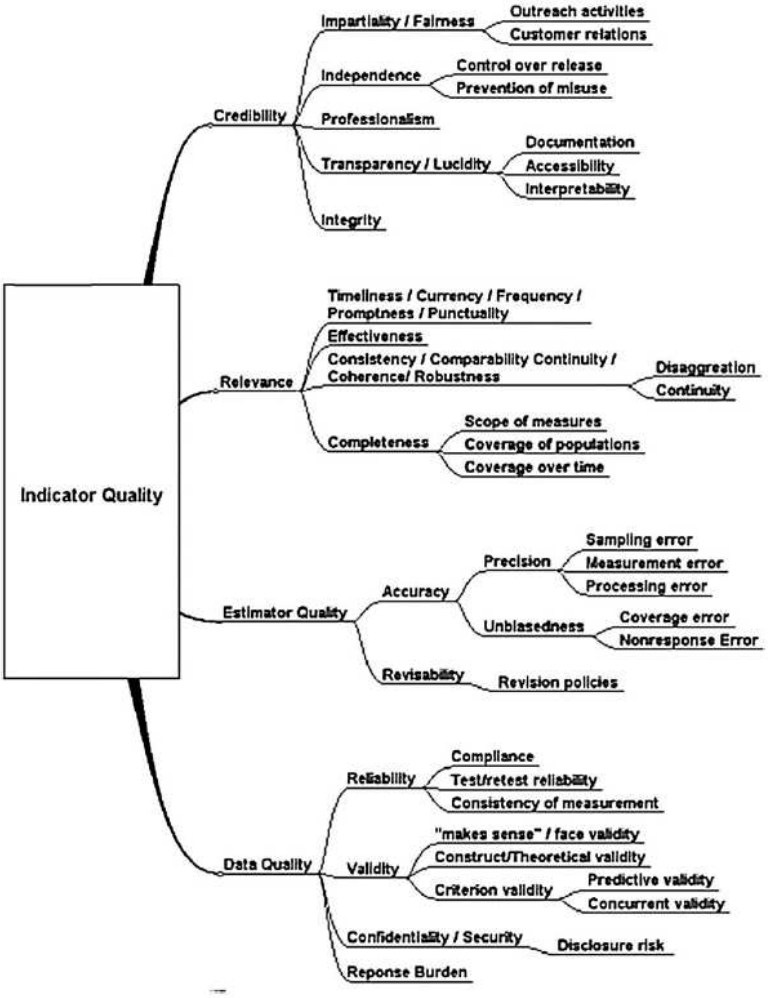


Figure 4. Draft Set of Indicator Quality Terms for Use by the Key National Indicators Initiative.