Survey analysis- week 2

Decomposing error and bias
"Total Survey Error"

© Peter Lugtig

p.lugtig@uu.nl

Class exercise: Your adopted survey

Form pairs and tell (and ask!):

1. What is the name of your survey?
2. What is the population of the study?
3. How are individuals selected to be invited for the survey? (sampling design)
4. What is the sample size?
5. What survey mode is being used?
6. What methods to prevent nonresponse?
7. What are the central concepts that are measured in the survey?

Your adopted survey

- Who thinks he/she has the smallest survey?
- Who has the largest survey?

- Who has a special population?
  – Not the general population

- What mode is used to interview people?
  – Face-to-face
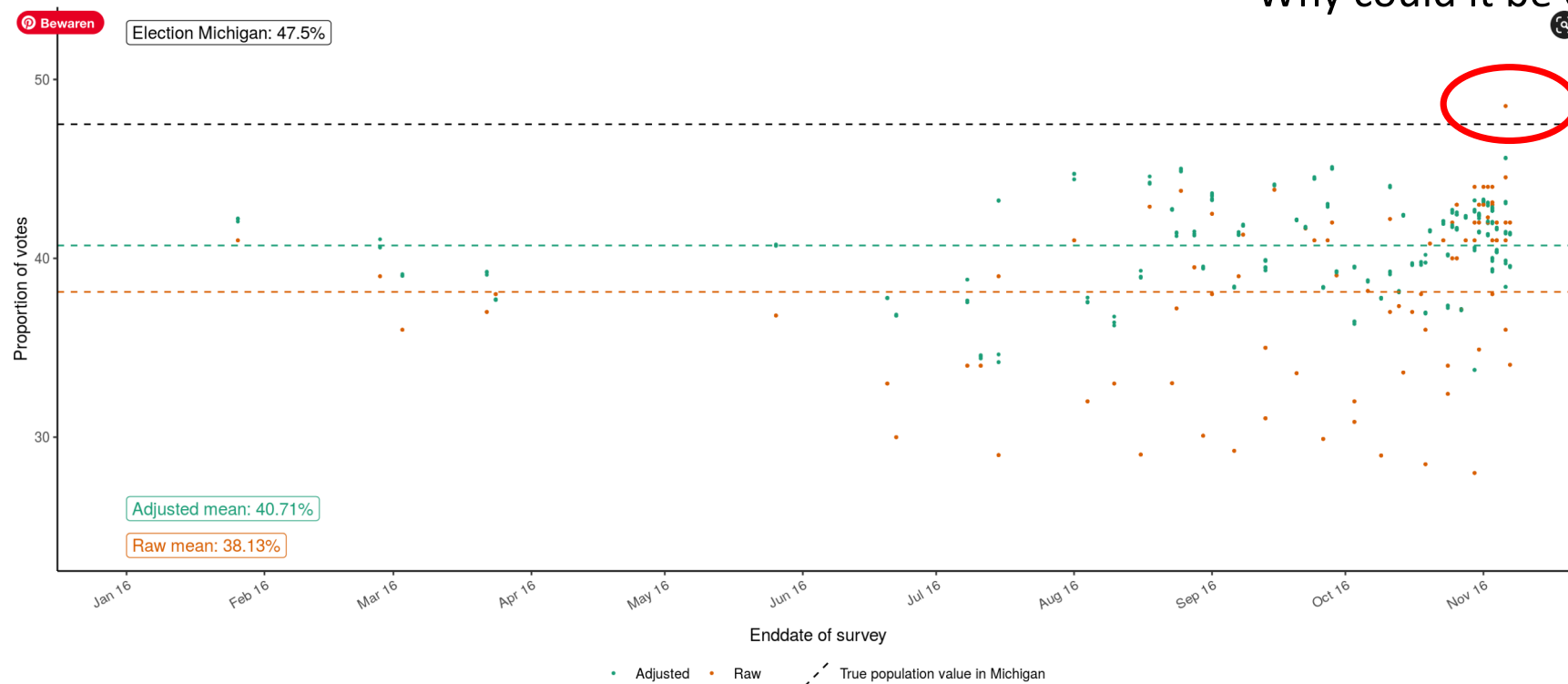 – Mail
  – Telephone
  – Web

Today

- Bias, error and MSE

- Total Survey Error (TSE)

- Total data error (TDE)

- How survey design affects error
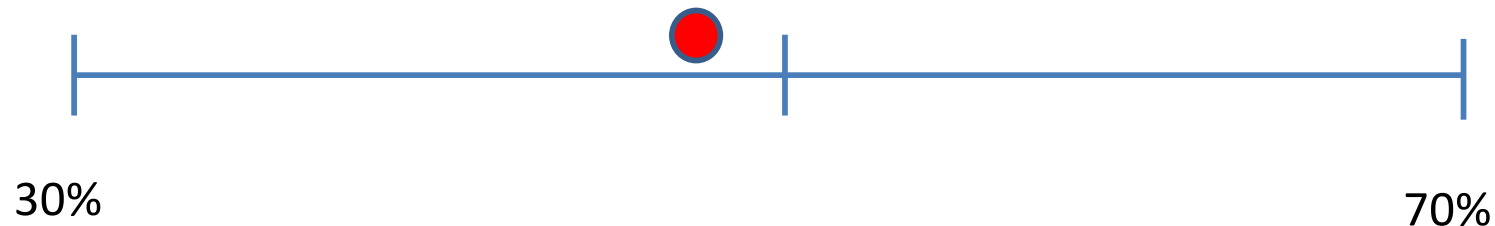 - The central role of survey modes

Class exercise week 1
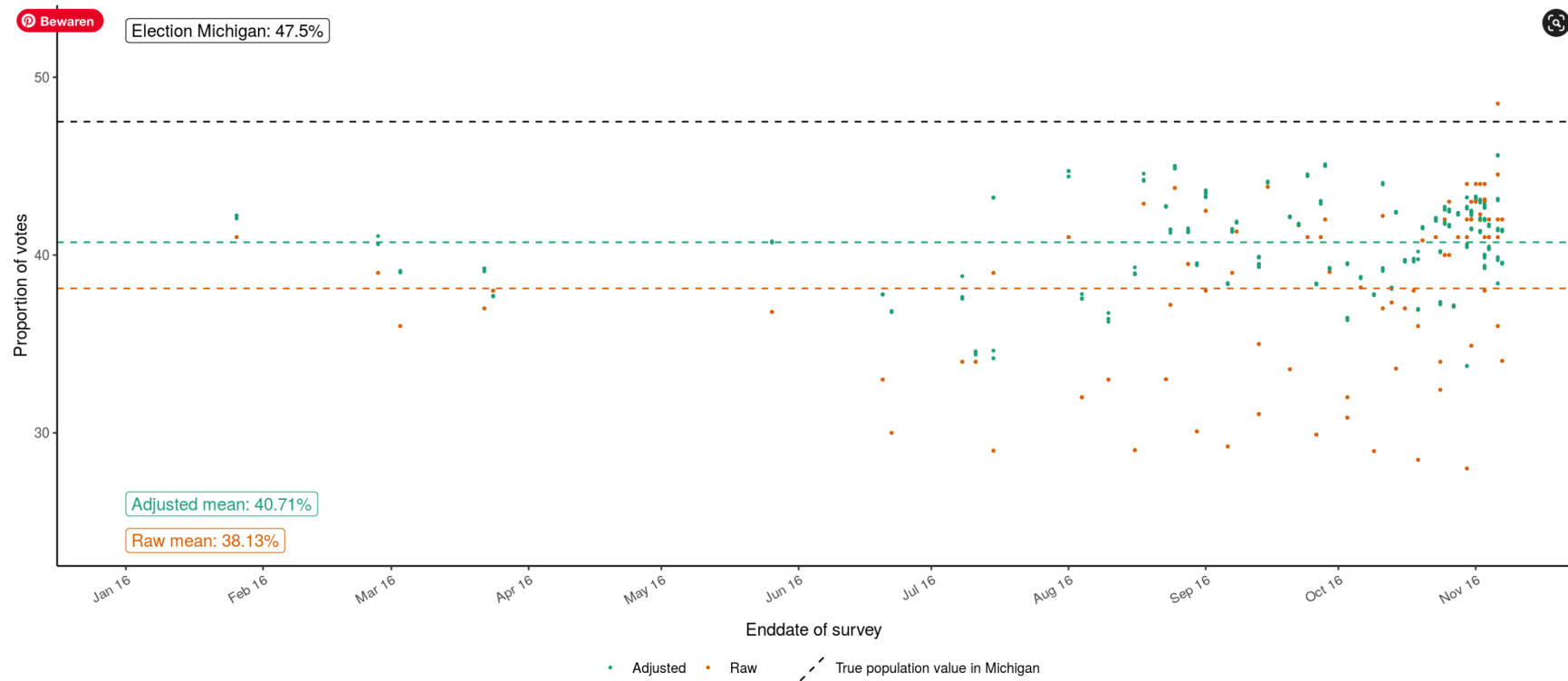
Class exercise week 1

Let's focus on 1 sample:
Why could it be off?
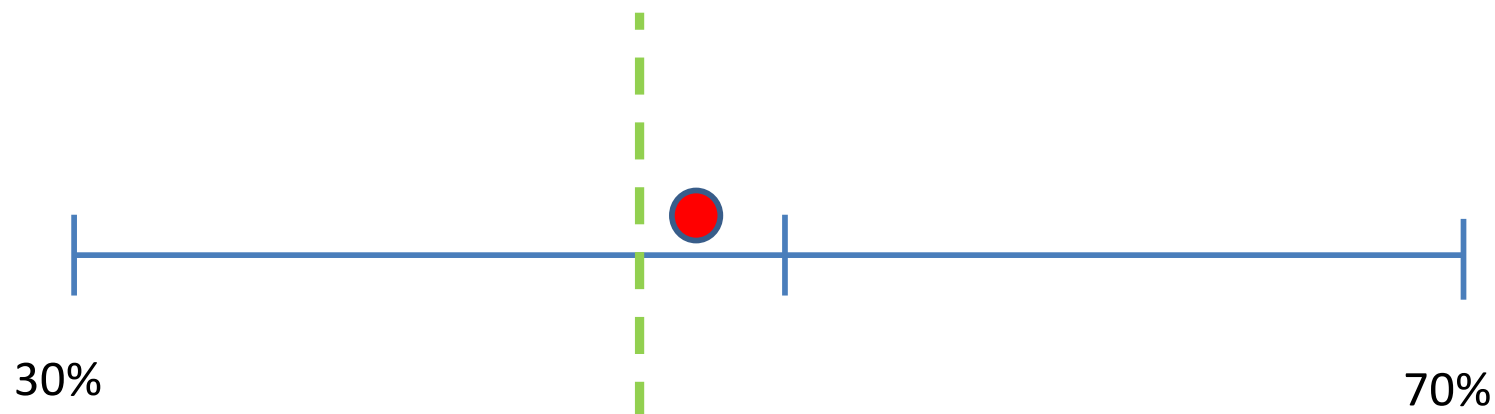
1 sample in Michigan: point estimate

- Estimate: 48%
- n= 1000

30%   70%

1 sample: bias

- Estimate: 48%
- True value: 47.5%
- Bias: $\hat{B} = \overline{y_r} - \bar{\mu}$ (Biemer, 2010)

  $\hat{B}$ = 48-47.5 = 0.5%

30% ———————————————— 70%

1 sample: error

- Point estimate: 48%
- Standard error (SE) = $\sqrt{p(1-p)/n}$

  SE = $\sqrt{(.48*.52)/1000}$ = .016
- Confidence interval: [.45 - .51]
  - [p +- 1.96 * se]
- Is error or bias a bigger problem?

30%   70%
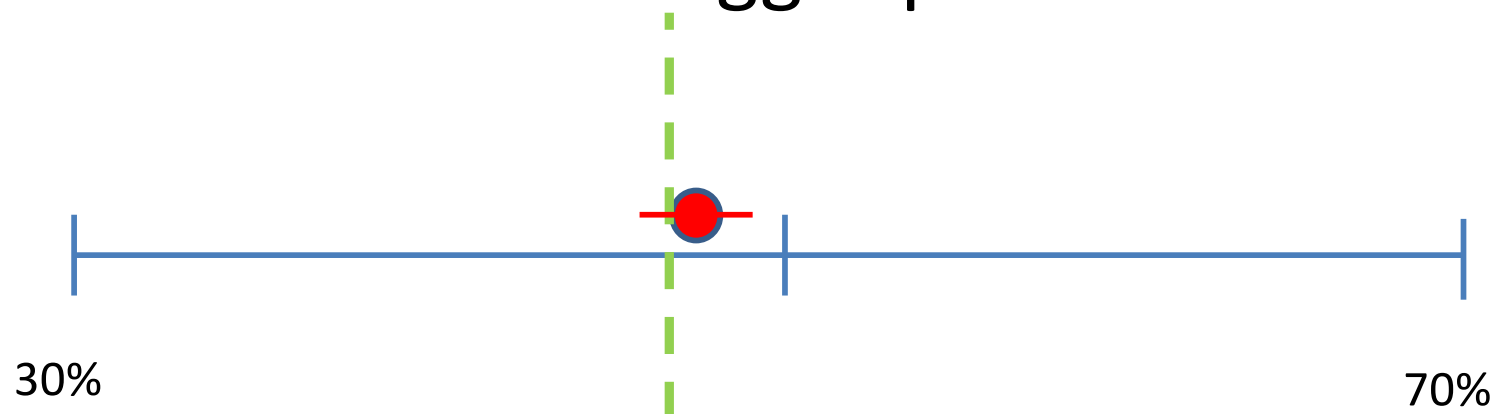
Estimate = true value + error + bias

- True value is within confidence interval!
- So, we have to <u>assume</u> there is bias
  - The 0.5% difference could be bias or error
    - Or, is the bias we see higher than expected error?
  - Next week, more on bias and error in samples
- Mean Squared error (MSE): $bias^2$ + error

30%   70%
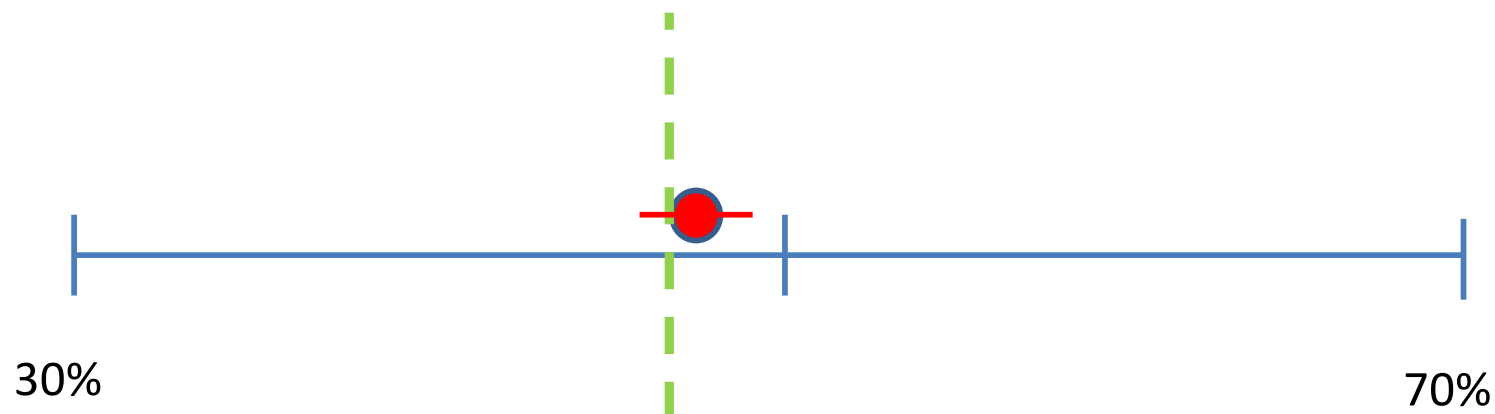
Mean square error

- MSE = $.005^2$ + .016 = .000025 + .016 = .016025

- Sampling error seems larger problem

- Biemer: a bit more complicated as true value also has a variance $\widehat{MSE}(\bar{y}_R) \doteq \widehat{B}^2 - v(\bar{\mu}) + 2\sqrt{v(\bar{y}_R)v(\bar{\mu})},$

  – Estimating population variance in mean tricky…

30%   70%
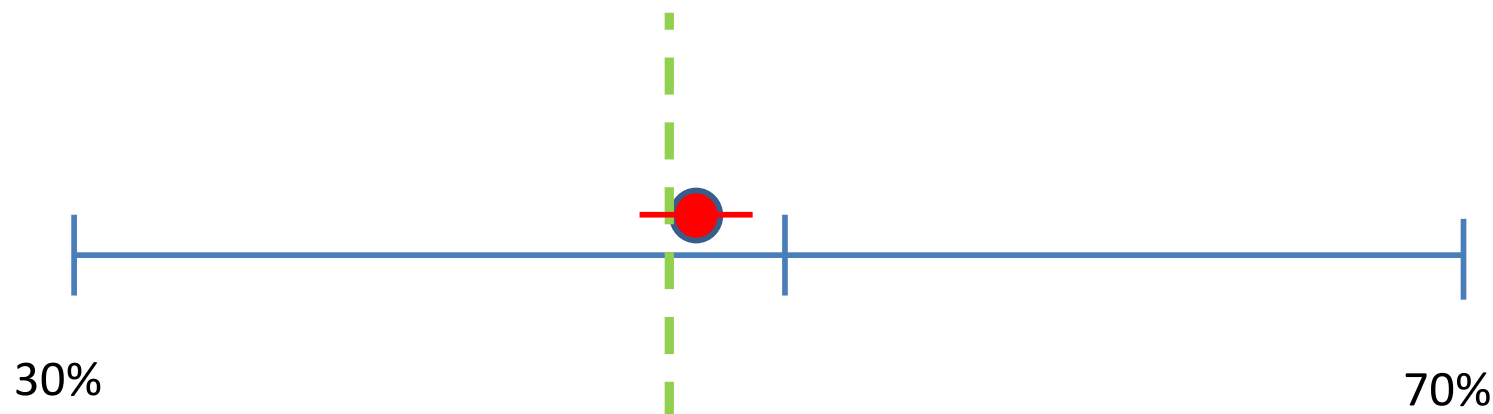
Survey design

- What if we increase sample size?
  - 10000 instead of 1000?
  - Se(r): $\sqrt{(.48*.52)/10000}$ = .005

Margin of Sampling Error at Specified Proportions

Assumptions: Simple random sampling with 95% confidence intervals

Proportion — 1% — 5% — 10% — 25% — 50%
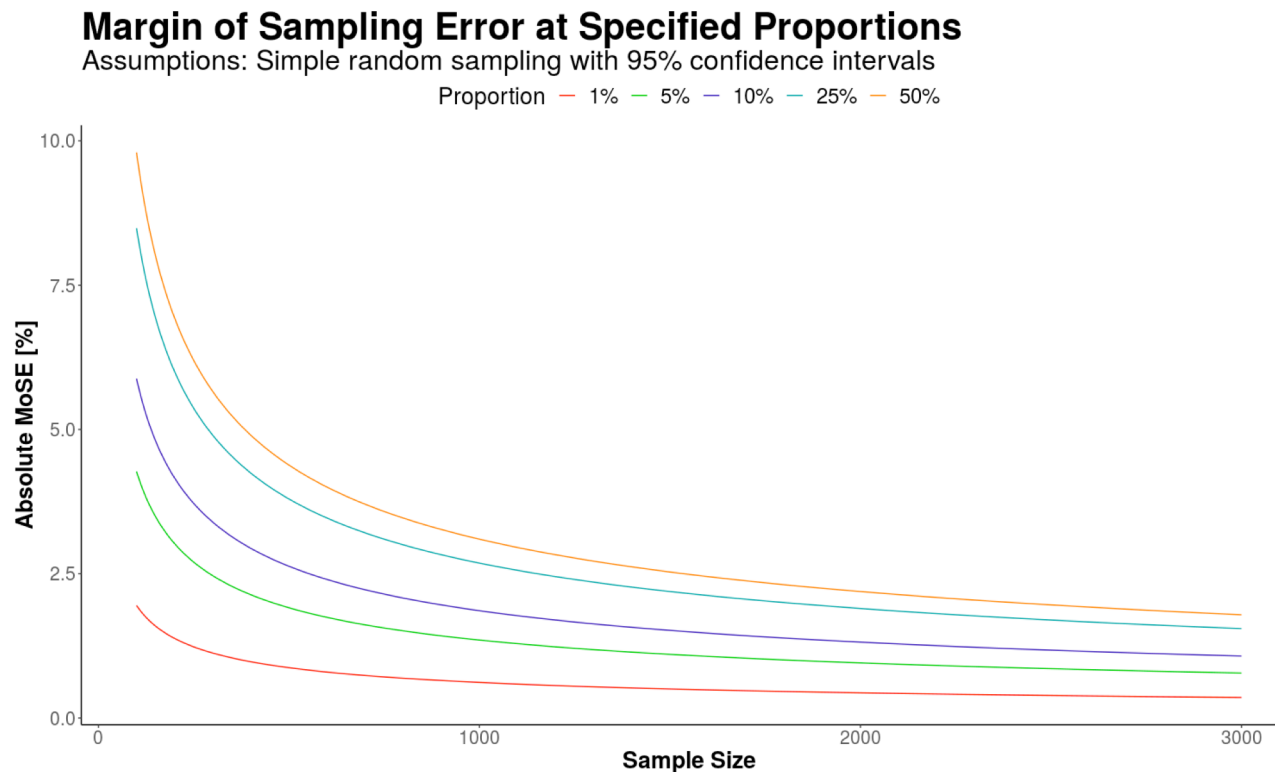
Why we still worry about bias
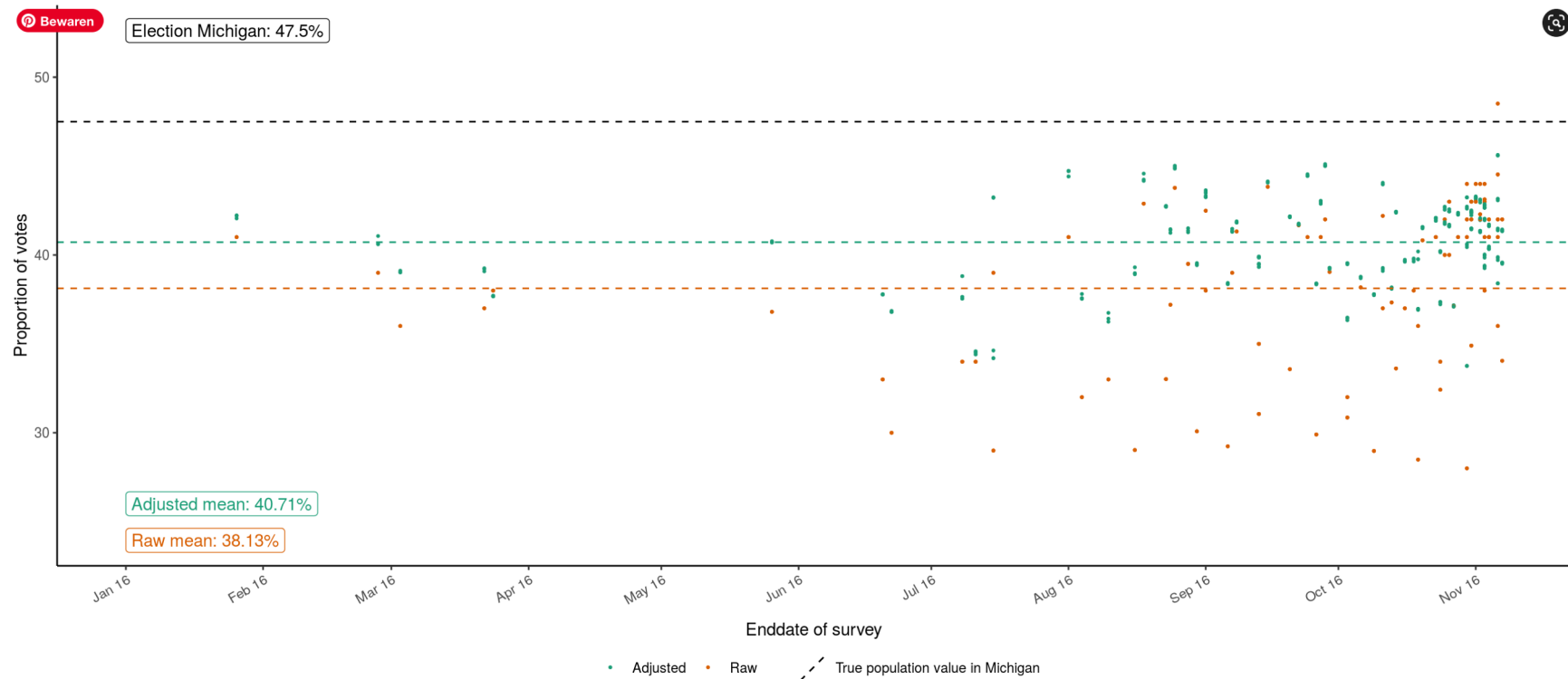
Bias and error in more detail

Total Survey Error

Total Survey Error Framework
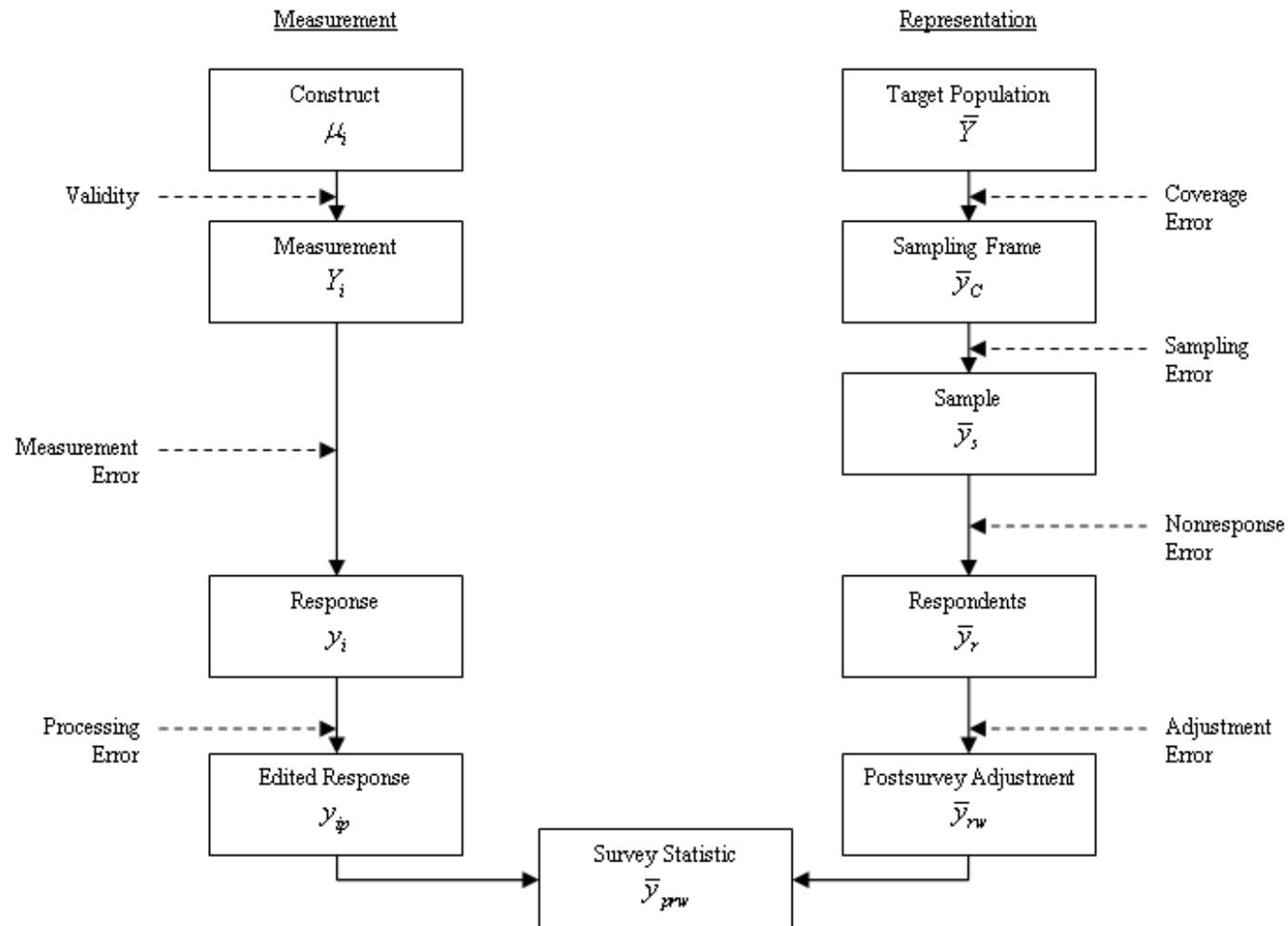
(Groves et al. 2009, p.48)

Total Survey Error (TSE) Framework

Measurement

Construct

Coverage Error

Discrepancy between target population and available sampling frames

Response
y_i

Edited Response
y_{ip}

Survey Statistic
$\bar{y}_{prw}$

Representation

Target Population
$\bar{Y}$

Coverage Error

Sampling Frame
$\bar{y}_C$

Sample
$\bar{y}_s$

Respondents
$\bar{y}_r$

Postsurvey Adjustment
$\bar{y}_{rw}$

(Groves et al. 2009, p.48)
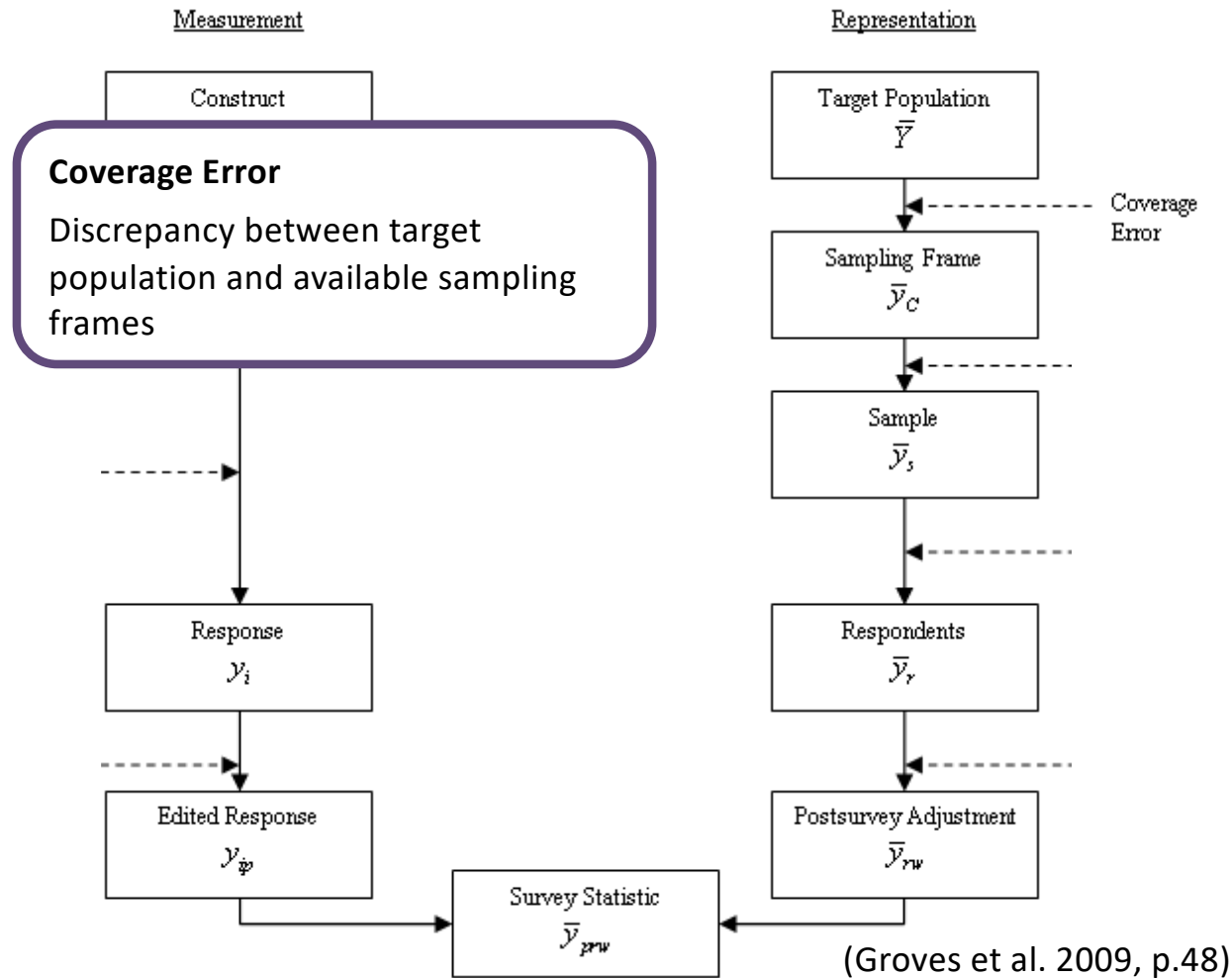
Terminology: coverage error

- (target) Population: group of units (people, companies, households, etc.) you are studying
- Sampling frame: list containing population elements
- Undercoverage: should be on, but is not
  - Not have an address, telephone, e-mail
- Overcoverage: should not be on, but is
  - Two phones, multiple e-mail, has died, has moved

- From: population register, schools, health records

Coverage error and modes

- Modes:
  - Web: no lists of e-mailadresses (unless special population) 🙁
  - Paper: invitations by <u>mail</u> to <u>households</u> 🙁
  - Face-to-face: Use list addresses or random walk 🙁
  - Telephone: Random Digit Dialing, mobile phones 🙁

- Problems with lists
  - Seldom up-to-date
  - Getting access is difficult

TSE – sampling error

Measurement

Construct
μ_i

Sampling Error

Originates from not observing all units in a sampling frame, but just a random sub-sample. This is why we calculate standard errors, confidence intervals etc.

Response
y_i

Edited Response
y_{ip}

Survey Statistic
$\bar{y}_{prw}$

Representation

Target Population
$\bar{Y}$

Coverage Error

Sampling Frame
$\bar{y}_C$

Sampling Error

Sample
$\bar{y}_s$

Respondents
$\bar{y}_r$

Postsurvey Adjustment
$\bar{y}_{rw}$

Terminology: sampling error

- Sampling unit: collection of units to be sampled from your frame
- Sample: the actual units you sample
- Respondents: the people out of the sample who participate

- Sampling can introduce bias and error!
  – Selecting people within households
  – Villages, hospitals, etc, etc.

TSE – nonresponse error

Measurement

Construct
μ_i

Measurement
Y_i

Nonresponse Error

The discrepancy between the gross sample and the net sample of observed unit.

y_i

Edited Response
y_{ip}

Survey Statistic
$\bar{y}_{prw}$

Representation

Target Population
$\bar{Y}$

Coverage Error

Sampling Frame
$\bar{y}_C$

Sampling Error

Sample
$\bar{y}_s$

Nonresponse Error

Respondents
$\bar{y}_r$

Postsurvey Adjustment
$\bar{y}_{rw}$

(Groves et al. 2009, p.48)

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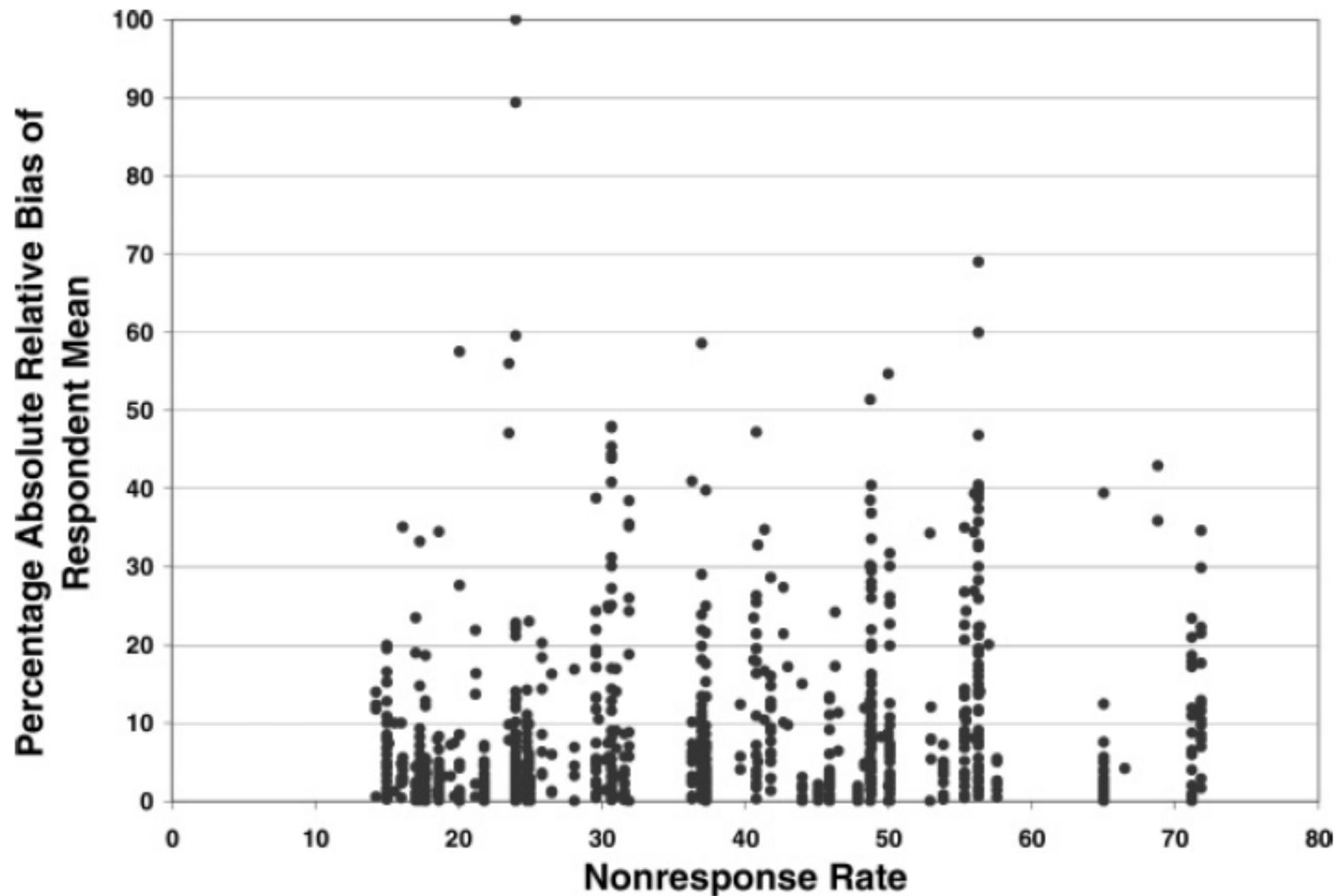
Figure 2. Percentage Absolute Relative Nonresponse Bias of 959 Respondent Means by Nonresponse Rate of the 59 Surveys in Which They Were Estimated.

Source: Groves, R. M., & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: a meta-analysis. *Public opinion quarterly*, *72*(2), 167-189.

TSE – adjustment error

Measurement

Construct
μ_i

Measurement
Y_i

Adjustment Error

By means of post-survey adjustments (e.g. weighting and imputations) we aim to correct for coverage, sampling and nonresponse errors.

Representation

Target Population
$\bar{Y}$

Sampling Frame
$\bar{y}_C$

Sample
$\bar{y}_s$

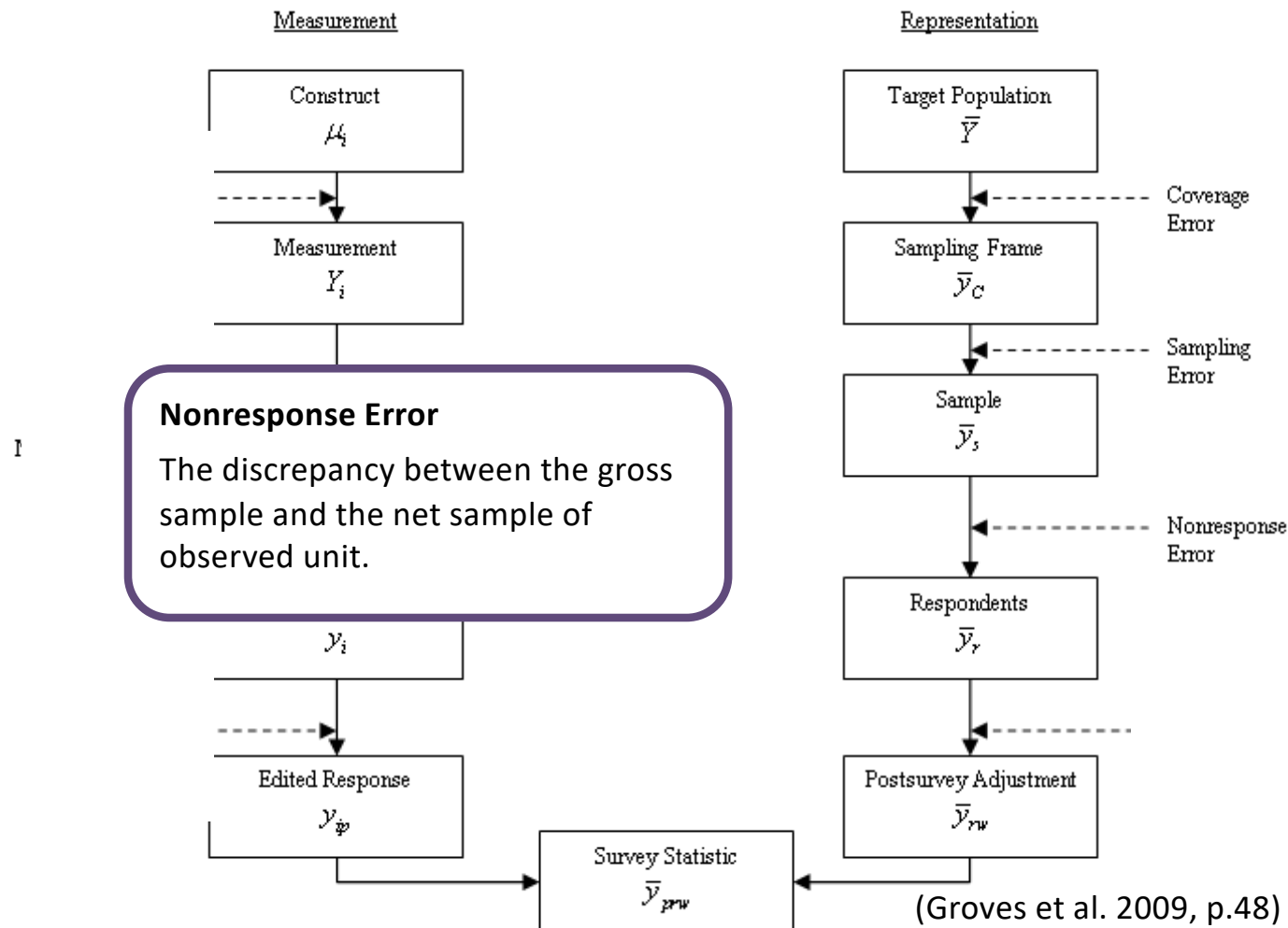Respondents
$\bar{y}_r$

Postsurvey Adjustment
$\bar{y}_{rw}$

Coverage Error

Sampling Error

Nonresponse Error

Adjustment Error

$\bar{y}_{prw}$

(Groves et al. 2009, p.48)

TSE – construct validity

(Groves et al. 2009, p.48)

TSE – measurement error

Measurement

Representation

Construct
μ_i

Validity

Measurement
Y_i

Measurement
Error

Response
y_i

Edited Response
y_{ip}

Target Population
$\bar{Y}$

Coverage
Error

Sampling Frame
$\bar{y}_C$

Measurement Error

The discrepancy between ideal measurement and actual answers given, e.g. because of:

• Misunderstandings

• Social desirability bias

• Satisficing

Adjustment
Error

Postsurvey Adjustment
$\bar{y}_{rw}$

Survey Statistic
$\bar{y}_{prw}$

(Groves et al. 2009, p.48)

TSE – processing error

(Groves et al. 2009, p.48)

In sum

- In design-based surveys:
  - Sampling error only error we know
  - We can control **by increasing sample size**
  - But many more sources of error/bias
    - Hard to always quantify exactly

- It is not strange polls are off!
- Key question in Survey design:
  - In order to minimize MSE:
  - Do we invest in larger sample, or more nonresponse follow-ups? Incentives, etc?

TSE and survey design

- Design aspects greatly affect survey errors
  - Invitation mode
  - Administration mode
  - Interviewers
  - Incentives, Questionnaire length
- Right design depends on:
  - Population
  - Topic of the study
  - Availability of sampling frames
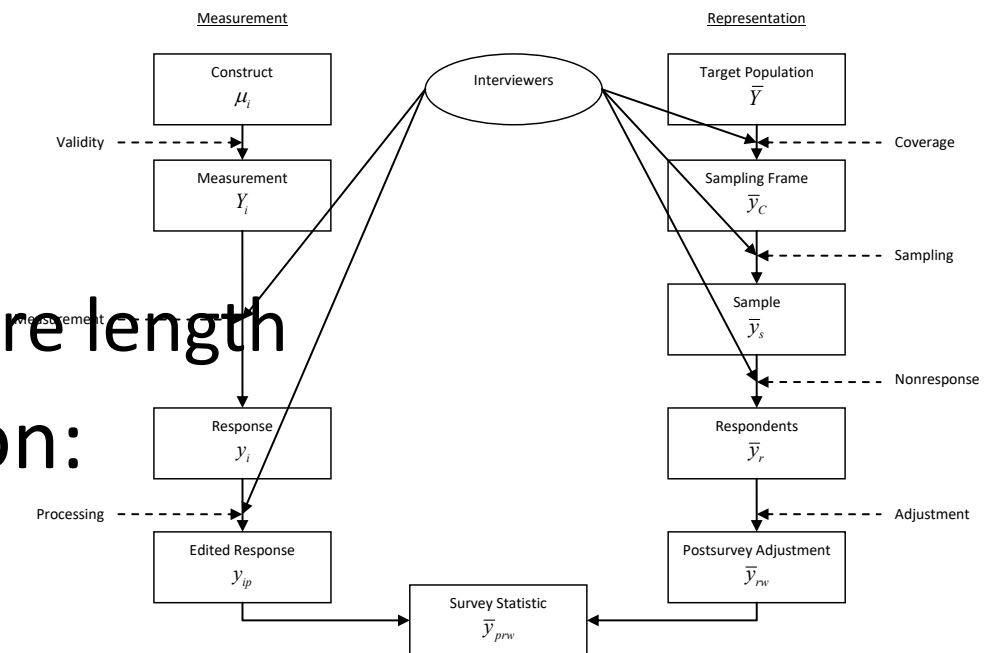
Groves and Lyberg (2010)

- History of the TSE framework

- There are extensions to
  - comparative surveys (Smith 2010),
  - longitudinal surveys (Lynn and Lugtig 2017),
  - analytic error (West and Saskhaug 2018)
  - … and Total Data error (other data sources)
    - Administrative, sensor, social media, etc.
    - See also weeks 13,14

TSE and model-based inference

- Convenience samples
- Volunteer opt-in panels
- (Quota samples)
- Lab-studies (psychology), organic data, social media, etc.

Model-based inference

Representation

Target Population
$\bar{Y}$

Coverage
Error

Sampling Frame
$\bar{y}_c$

Sampling
Error

Sample
$\bar{y}_s$

Nonresponse
Error

Respondents
$\bar{y}_r$

Adjustment
Error

Postsurvey Adjustment
$\bar{y}_{rw}$

Survey Statistic
$\bar{y}_{prw}$

Compare data to population, model **selection bias**
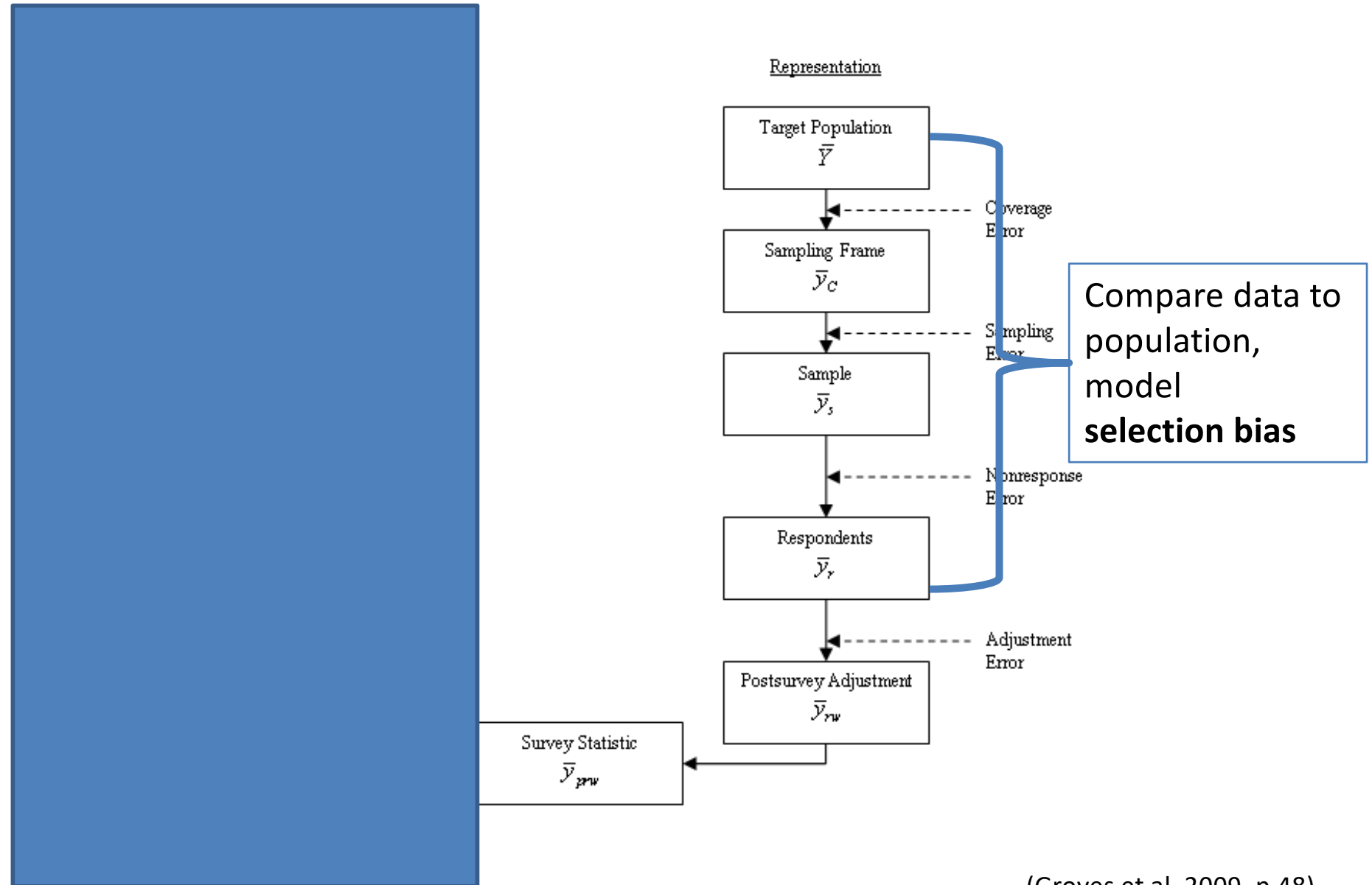
(Groves et al. 2009, p.48)

Class Exercise (and then break)

Form groups of 4 and discuss for your scenario:

- What component of TSE are you worried about?

- What can you do in design to minimize MSE
    - Think about error and bias
        - Minimize error by increasing sample size
        - Minimize bias by thinking about measurement, sampling, nonresponse

- 15 min + 10 minutes break

Class exercise – worried about

	1. Flats in Utrecht	2. Homosexual muslims	3. EU-wide election study	4. (long)- Covid in Italy
Coverage				
Sampling				
Nonresponse				
Adjustment				
Validity				
Measurement				
Processing				
	Design:	Design:	Design:	Design:

In sum: How much to worry about each?

	Bias	Error
Coverage error	**	
Sampling error	*	***
Nonresponse error	***	*
Adjustment error		*
Validity of measurement	***	
Measurement error	**	**
Processing error		*

Class Exercise: group 1

- A researcher would like to know to what extent neighbours in high-rise flats (over 8 floors tall) in Utrecht help eachother out. She suspects that people help eachother mainly if they have the same ethnic and socio-economic background.

Class Exercise: group 2

- A researcher would like to do a survey among homosexual muslims in the region of Utrecht to find out how the families of these men and women deal with this.

Class Exercise: group 3

- For the next elections for the European parliament, a market research firm with offices in all EU countries would like to do a pan-EU survey among the EU electorate to

- a) predict the outcome of the election in every country and

- b) compare the attitudes of people in different countries towards the European Parliament.

Class Exercise: group 4

- A researcher would like to better understand how patients who developed Covid-19 in the spring of 2020 (March-april) in Italy are now recovering from their illness. There is no central registry of patients in Italy; these are kept at hospitals, and if you want to reach these patients it is necessary to collaborate with individual hospitals in Italy.

- In the survey you want to ask questions about phsyical and mental wellbeing, as well as the effects Covid-19 has had on relations with household members (children, partner).

Next week

- Prepare
  - 1. Read Stuart (see e-mail last week)
    - Simple Random sampling
  - 2. Do THE exercise 2 (sampling)
- Lecture on Simple Random Samples