Evaluating and reducing biases in mixed-mode survey data

AAPOR Webinar November 15 2018

Thomas Klausch (t.klausch@vumc.nl)
Amsterdam University Medical Centers

Barry Schouten (jg.schouten@cbs.nl)
Statistics Netherlands / Utrecht University

Introduction

A mixed-mode survey combines multiple modes of administration in the same design

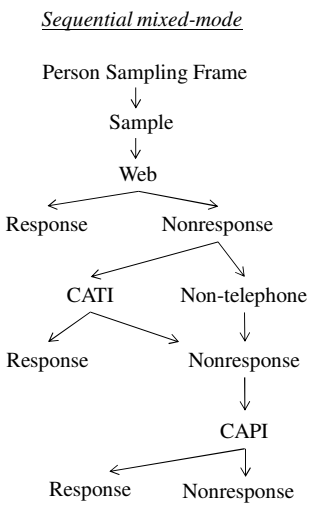
Objectives of mixed-mode surveys

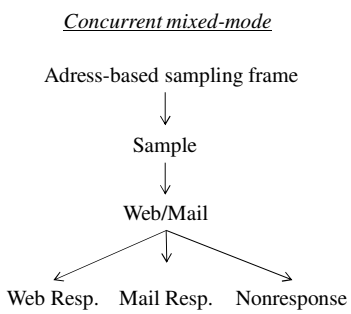
- Save costs
- Increase response rates compared to single-mode design
- Reduce non-response and coverage bias

Estimates from mode-specific respondent sets often differ due to

- Selection effects (= different types of respondents, in expectation)
- Measurement effects (= different answers given at same 'true state' , in expectation)
- Total effects (= the sum of both, a realization is observable in mixed-mode data)

The "dominant design" types: sequential and concurrent





Examples (sequential):
American Community Survey
Crime Victimization Survey (NL & others)

Overview on Today's Webinar

1. **Defining and understanding measurement and selection effects**
   - Using true scores / gold standard
   - Using potential outcomes framework
2. **Covariate-based estimation of measurement and selection effects**
   - Class of techniques exploiting missing at random (unconfoundedness) assumptions
   - Exogenous information (e.g. from sampling frame)
   - Regression, weighting, (multiple) imputation
3. **Adjusting mode effects using alternative designs and assumptions**
   - Instrumental variables
   - Re-interview data
   - Time-series stabilization

Part 1:

Defining and understanding measurement and selection effects

Defining selection effects as contrast of selection biases.

Let y_i be a true score for unit i and r_i^a an indicator denoting response in mode a.

Let $\hat{\mu}_a$ denote the true score mean for respondents in mode a, i.e.

$$\hat{\mu}_a = \frac{1}{n_a} \sum_i r_i^a y_i \, .$$

with expectation $E_r[\hat{\mu}_a] := \mu_a$. The selection bias of mode a is

$$SB_a = \mu_a - \mu \, .$$

with μ the population mean. The selection effect between modes a and b is

$$SE = SB_b - SB_a = \mu_b - \mu_a \, .$$

Defining measurement effects as contrast of measurement biases.

Let ϵ_i^a be the measurement error of mode a. The measured outcome of mode a is

$$\mathbf{y_i^a = f_a(y_i, \epsilon_i^a)\ , for\ example\ y_i^a = y_i + \epsilon_i^a\ .}$$

So response y_i^a is a function of its true score. The observed respondent mean in mode a is

$$\mathbf{\hat{\mu}_a^a = \frac{1}{n_a} \sum_i r_i^a y_i^a}$$

with expectation $E_{r,\epsilon}[\hat{\mu}_a^a] := \mu_a^a$. The measurement (error) bias of the respondent set in mode a

$$\mathbf{MB_a = \mu_a^a - \mu_a.}$$

Defining measurement effects as contrast of measurement biases.

The measurement effect between modes a and b is

$$\mathbf{ME = MB_b - MB_a = (\mu_b^b - \mu_b) - (\mu_a^a - \mu_a).}$$

The total effect between the modes is

$$\mathbf{TE = \mu_b^b - \mu_a^a\ .}$$

Now note that

$$\mathbf{TE = ME + SE\ .}$$

We are often interested in decomposing the total effect (= estimating its components).

An example using record checks approach:

A **web-telephone mixed-mode survey** in which respondents are asked for the gross monthly income. For each respondent in the sample you know his **true income** from the tax office registry.

You use the register to determine for web respondents the true mean income at 3500$. For telephone respondents it is 2800$.

The mean answers of web respondents is 3600$. For telephone respondents we have 3200$.

We have:

$\hat{\mu}_{web} = 3500, \hat{\mu}_{tel} = 2800$, $\hat{\mu}_{web}^{web} = 3600, \hat{\mu}_{tel}^{tel} = 3200$

$\widehat{SE} = 2800 - 3500 = -700$; $\widehat{ME} = 400 - 100 = 300$; $\widehat{TE} = 3200 - 3600 = -400$

The bias of the mixed-mode estimate

The pooled mixed-mode respondent mean of a design with modes a and b is

$$\hat{\mu}_{m}^{m} = \frac{1}{n}\sum_{i} (r_i^a y_i^a + (1 - r_i^a) y_i^b)$$

Proportion of mixed-mode respondents $\pi_a := E_r[r_i^a]$. So

$$TB_m = \pi_a(SB_a + MB_a) + (1 - \pi_a)(SB_b + MB_b)$$

So that

$$SB_m = \pi_a SB_a + (1 - \pi_a) SB_b$$

and MB_m similarly (replace SB by MB above). And of course

$$TB_m = SB_m + MB_m .$$

An example using record checks approach:

For the web-telephone mixed-mode design we had:

$\hat{\mu}_{\text{web}} = 3500$, $\hat{\mu}_{\text{tel}} = 2800$, $\hat{\mu}^{\text{web}}_{\text{web}} = 3600$, $\hat{\mu}^{\text{tel}}_{\text{tel}} = 3200$

Now assume that we draw a sample of n=1000, of which 200 respond in web, 300 respond in telephone. The response rate is 50% with

$$\hat{\pi}_{\text{web}} = \frac{2}{5} = 40\%$$

Assume the true population mean is 3300$. Then:

$\hat{\mu}^{\text{m}}_{\text{m}} = \frac{2}{5}3600 + \frac{3}{5}3200 = 3360$ with $\widehat{\text{SB}}_m = -220$, $\widehat{\text{MB}}_m = 280$, and $\widehat{\text{TB}}_m = 60$.

What is the impact of selection and measurement effects on mixed-mode total bias?

Complex interplay between effects and biases (sometimes called "mix")

$$\mathbf{TB_m = [SB_a + (1 - \pi_a)SE] + [MB_a + (1 - \pi_a)ME]} \text{ .}$$

In the example: $\widehat{\text{SB}}_{\text{web}} = 200$, $\widehat{\text{MB}}_{\text{web}} = 100$, $\widehat{\text{SE}} = -700$, $\widehat{\text{ME}} = 300$, $\hat{\pi}_{\text{web}} = \frac{2}{5}$

$$\mathbf{\widehat{TB}_m = \left[200 + \frac{3}{5}(-700)\right] + \left[100 + \frac{3}{5}300\right] = 60}$$

Participant poll – mode effect estimation

Does mode effect estimation play a role at your institution?

- Yes
- No
- Don't know

How can we estimate biases and mode effects in practice?

Record check approach

- True scores available from an external source
- All biases and effects can be estimated
- Not feasible in practice

Measurement benchmark mode approach

- Define reference mode to produce best answers for a question
- Set benchmark measurements equal to true scores
- Choose by methodological argument (e.g. social desirability low in web) or
- Ex-post empirical argument (e.g. choose mode with less desirable answers as benchmark)

Introduction of benchmark mode requires potential outcomes

Mode a is benchmark: $y_i^a = y_i$ and

$$\mu_a = \boldsymbol{\mu_a^a}$$

which means that there is no MB of mode a,

$$MB_a = \mu_a^a - \mu_a = \mathbf{0}\ .$$

For mode b we now have

$$\mu_{\mathbf{b}} = \mu_{\mathbf{b}}^{\mathbf{a}} \coloneqq \mathbf{E}_{\mathbf{r},\epsilon}\left[\frac{\mathbf{1}}{n_b}\sum_i r_i^b y_i^a\right]$$

where μ_b^a is the mean of the answers that mode b respondents potentially would have given under mode a (= potential outcomes, Rubin 1974)

Consequences for mode effects

Selection effects in potential outcome notation:

$$\mathbf{SE} = \boldsymbol{\mu_b - \mu_a} = \boldsymbol{\mu_b^a - \mu_a^a}$$

Measurement effects:

$$\begin{aligned}\mathbf{ME} &= \left(\boldsymbol{\mu_b^b - \mu_b}\right) - \left(\boldsymbol{\mu_a^a - \mu_a}\right)\\ &= \left(\boldsymbol{\mu_b^b - \mu_b^a}\right) - \left(\boldsymbol{\mu_a^a - \mu_a^a}\right)\\ &= \boldsymbol{\mu_b^b - \mu_b^a}\end{aligned}$$

As before **TE =ME+SE.**

To estimate the mode effects we need an estimate of the potential outcome mean μ_b^a.

Impact of mode effects on mixed-mode bias under benchmark approach

Total bias of the mixed-mode respondent mean is now

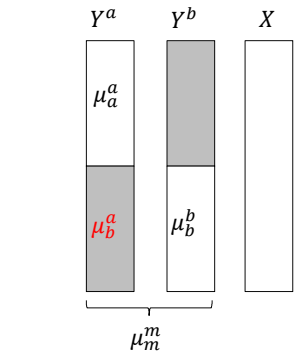
$$\mathbf{TB_m = [SB_a + (1 - \pi_a)SE] + (1 - \pi_a)ME}$$

1. Reduce selection bias of mode a, if SE has alternate sign of SB
2. The ME can increase the TB, especially if the SB_a is nullified by the SE
   - Offsets the gain from the SE
   - Means mode b gives different answers
3. A ME with appropriate sign can still reduce TB
   - As we usually do not have knowledge of size of SB_a, this possibility is not considered
   - Instead: want to **have SE**, want to **avoid, estimate, and adjust ME**

Part 2:

Covariate-based estimation of measurement and selection effects

Overview: estimating potential outcomes is a missing data problem.



Observed (Response)

Unobserved (potential outcomes)

Missing data problem akin treatment effect estimation in observational studies.

Same techniques used

- Regression estimation
- Weighting
- Combinations (double-robust estimation)
- Multiple imputation

Same assumption

- Missing at random potential outcomes
- Exogeneity of auxiliary data

Auxiliary data (covariates) and exogeneity assumption

Sources of auxiliary data:

- Sampling frame
- Survey questions
- Paradata (survey / process data)
- Panels

Have to be "exogenous":

- Cannot be impacted by measurement effects; untestable
- Assured for sampling frame data
- Not given for many survey questions (exception: simple factual questions, e.g. sex)
- Difficult to get mixed-mode paradata free of MEs

Auxiliary data (covariates) and missing at random assumption

Missing at random (MAR) potential outcomes (unconfoundedness)

- Conditional on X the potential outcomes Y^a and Y^b are independent of response indicator R

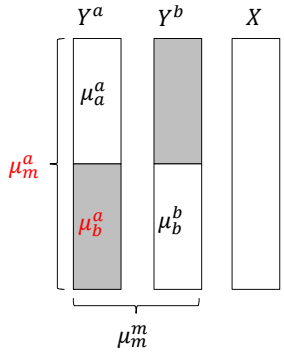
When given?

- X are the common causes of mode-specific response and the potential outcomes
- Techniques requires strong predictors to work well

Problem in practice

- Is the MAR assumption plausible given the (few) exogenous variables available?
- **Problem**: survey with sampling frame age, gender, household size – does this justify MAR?
- **Potential**: panel surveys with baseline single-mode recruitment interviews

Covariate-based potential outcome estimation in a nut-shell.



Observed (Response)

Unobserved (potential outcomes)

Regression estimation

- Model: $\mu_a^a = \mathbf{X\beta}$
- Estimate: $\hat{\mu}_b^a = \frac{1}{n_b}\sum_i \mathbf{r_i^b} X_i \hat{\beta}$

Weighting estimation

- Model: $\Pr(R = b|X) = \text{logit}(\mathbf{X\beta}) = \rho$
- Estimate: $\hat{\mu}_m^a = \sum_i \mathbf{r_i^a y_i^a} \hat{w}_i$
- Weights: $\hat{w}_i = \hat{\rho}_i / \sum_i \hat{\rho}_i$

Variance estimation uses bootstrap

Other options: double robust, multiple imputation, matching

Part 3:

Adjusting mode effects using alternative designs and assumptions

Mode effect reduction and adjustment – getting started

Four decisions

- What is the objective, reduction or adjustment?
- Adjustment to what, i.e. choice of benchmark mode or design?
- What kind of auxiliary information is collected?
- What kind of trade off (bias, accuracy, MSE, comparability versus cost)?

Side remarks:

- Also single mode designs have mode effects
- Mode effect is a fuzzy term for multi-mode survey designs

Mode effect – reduction or adjustment?

- **Reduction (in design stage)**
  - Mode effect components are estimated but used only to (de)select modes for future rounds
  - Resulting survey designs may be adaptive mixed-mode when this decision is made differently for different (relevant) population subgroups
  - Implies a pilot or experimental study, and, hence, an investment at the design stage
- **Adjustment (in estimation stage)**
  - Mode effect components are estimated and used to correct survey estimates towards the specified benchmark design
  - Implies a cost-benefit analysis
  - Correction may assume time-stability of measurement effect and require assessment of mode effects at a low time frequency
  - Correction may also assume a time-varying measurement effect and demand for continuous assessment of mode effects

Participant poll – mode effect adjustment

Do you perform some form of mode effect adjustment for one or more of your surveys?

- Yes, to all surveys
- Yes, to some of the surveys
- No
- Don't know

Alternative choices of benchmark

Benchmark estimates have two components:

1. The measurement benchmark mode (introduced in part 2)
2. The selection benchmark mode/design

In part 2: mode a was measurement benchmark and we chose no selection benchmark.

Single mode design benchmark

- E.g. F2F measurements and F2F selection deemed best)

Hybrid benchmark designs

- E.g. Selection of a sequential web-F2F design and the measurement of F2F are deemed best

Benchmark designs and time series

- Repeated versus non-repeated surveys

Auxiliary information

In order to break confounding of selection and measurement auxiliary information is imperative

- Selection effect viewpoint
  - Selection effect estimated and remainder is measurement effect
  - Uses administrative data, paradata and/or re-interview data
  - Part 2 discusses covariate-based estimation with standard administrative data and paradata
- Measurement effect viewpoint
  - Measurement effect is estimated and remainder is selection effect
  - Uses paradata, latent constructs and/or re-interview data
- Randomization = instrumental variable
  - Selection neutralized by an instrumental variable, i.e. not related to measurement

Mode effect estimation and adjustment – literature

- Selection effect viewpoint
  - Suzer-Gurtekin (2013), Kolenikov & Kennedy (2014, JSSAM), Vannieuwenhuyze, Loosveldt & Molenberghs (2014, JOS), Buelens & Van den Brakel (2015, SMR), Park, Kim & Park (2016, AoAS), Fessler, Kasy & Lindner (2018, JEI)
  - Re-interview: Biemer (2001, JOS), Schouten et al (2013, SSR)
- Measurement effect viewpoint
  - Klausch, Hox & Schouten (2013, SMR), Cernat (2015, SRM), Cernat, Couper & Ofstedal (2016, JSSAM), Mariano & Elliott (2017, JSSAM)
  - Re-interview: Klausch et al (2017, JSSAM)
- Instrumental variable
  - Vannieuwenhuyze, Loosveldt & Molenberghs (2010, POQ)

Mode effect stabilization in time

Simplest form of adjustment uses only the paradata on mode of response

- Idea: Calibrate response to fixed mode contributions (Buelens & Van den Brakel, 2015):
  - Set a mode response distribution, e.g. 50% Web, 25% telephone and 25% face-to-face, based on historic time series. This is the benchmark;
  - Estimate shares of modes to response in a particular round, say 45%, 20%, 35%;
  - Adjust respondent weights (in addition to regular weighting), i.e. 10/9, 5/4 and 5/7;
- Applicable only when:
  1. mode contributions do not fluctuate very strongly;
  2. mode contributions do not have trends.
- Assumptions:
  - Nonresponse and measurement error are independent
  - Time-stability of measurement error

Instrumental variable approach

Mixed-mode design is run parallel to a single mode design with same representation

- Single mode design is treated as benchmark and a two mode sequential design is adjusted
- Assumption:
  - Selection bias of the mixed-mode design is the same as that of single-mode design (representativity assumption), possibly after applying calibration or matching
- Drawback:
  - Strong representativity assumption
  - Approach cannot easily be extended to designs with >2 modes
- Example: Political interest

  Vannieuwenhuyze et al (2010, POQ)

ME effect	Size	Standard error	p
not at all	0,005	0,021	0,823
hardly	0.093	0,037	0,012
quite	-0,094	0,041	0,023
very	-0,004	0,025	0,877
SE effect			
not at all	-0,046	0,028	0,100
hardly	-0,049	0,060	0,420
quite	0,097	0,072	0,178
very	-0,002	0,046	0,964

SE estimation with re-interview

Employ a re-interview to respondents

- Starting point is again a sequential two mode design, but >2 modes is also possible
- The response each mode is calibrated to the combined response of the re-interview and follow-up.
- The remaining difference between modes is the ME and is removed;
- Assumptions:
  - SE can be estimated using mix of re-interview data, administrative data and paradata
  - Re-interview does not affect measurement behaviour of respondent
  - Nonresponse to re-interview is unrelated to survey variables of interest given administrative data and paradata

Re-interview Crime Victimization Survey

Dutch CVS re-interview with F2F follow-up to web, mail and telephone respondents

Web, mail and telephone response calibrated to combined re-interview and follow-up F2F response.

Adjustments:

Being a victim

	Coverage	Nonresponse	Measurement	Mode effect
Telephone	0,0%	0,2%	-4,0%*	-3,8%*
Paper	-	-1,6%	3,3%	1,7%
Web	1,3%	0,3%	3,9%	5,6%*

Number of victimizations

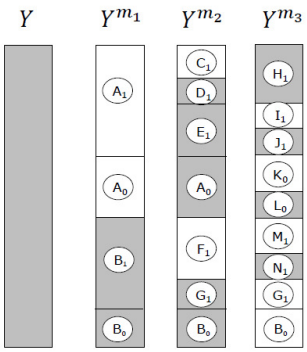
	Coverage	Nonresponse	Measurement	Mode effect
Telephone	0,3	-1,9	-4,7	-6,3
Paper	-	-3,3	12,5*	9,2*
Web	2,6	-3,3	15,3*	14,5**

Feeling unsafe at times

	Coverage	Nonresponse	Measurement	Mode effect
Telephone	-0,2%	-1,1%	-2,8%	-4,1%*
Paper	-	-0,2%	1,2%	1,1%
Web	0,4%	-0,6%	6,3%**	6,1%**

Re-interview design for three modes

Design of a re-interview for three modes



m1=web, m2=tel, m3=F2F

Dutch LFS: A_1, A_0, F_1, B_0 observed by default

American Community Survey: G_1 is also observed

X_1 Listed phone

X_0 No listed phone

Response

Potential outcomes

ME effect estimation with re-interview

Employ a re-interview to respondents

- Starting point is again a sequential two mode design, but >2 modes is again possible
- The selection of the mixed-mode design is treated as the benchmark and one of the modes is treated as benchmark for measurement:
  - The measurements in different modes are directly compared, possibly allowing for a rescaling: $y_{a,i} = \Delta + \lambda y_{b,i} + \epsilon_i$
  - The observed differences/fitted model are used to predict the potential outcomes;
- Assumptions:
  - Measurement equivalence: re-interview respondent answers are unaffected by first interview
  - Measurement difference between modes is independent of selection into re-interview
- Klausch et al (2017. JSSAM) compare properties of a range of estimators. And recommend the inverse regression estimator

Comparison of estimation and adjustment methods

Method	Data requirements	Assumptions	Advantages (+) Disadvantages (-)
Standard Covariate-based adjustment	• Sampling frame data • Paradata • Survey responses	• Exogeneity • MAR	• Too strong assumptions in many settings (-) • Adjustment on individual level possible (+)
Time-series stabilization	• Repeated cross-sectional / longitudinal survey	• Independence of meas. and sel. error • Time-stability of MEs	• Does not decompose (-) • Avoids ME estimation problem (+)
Instrumental variable method	• Single-mode reference survey parallel to mixed-mode	• Single-mode and mixed-mode survey have same SB	• Avoids MAR and exogeneity assumption (+) • Representativeness assumption usually implausible (-) • Not available for >2 modes
Re-interview method	• Re-interview of subset of mixed-mode respondents	• Measurement equivalence	• More plausible MAR assumption (+) • MNAR estimators available (+) (Klausch, 2017, JSSAM) • Measurement equivalence traded off against true score time-stability (-)

Extensions: re-interview cost-benefit analysis

Re-interview may be beneficial under three conditions

1. Survey variables of interest are relatively stable in time, i.e. high correlation between two measurements
2. Re-interview is likely not to affect respondent behavior
3. Measurement biases are anticipated/conjectured to be fairly large

- Viewpoints:
  - MSE(adjust) < MSE(unadjust), i.e. gain in bias outweighs loss in precision under fixed budget?
  - Precision maintained, i.e. var(adjust) = var(unadjust). How much extra budget required?
- Investment depends on:
  - Number of years/months mixed-mode design is kept fixed
  - Stable measurement bias assumption.

Extensions: cost-benefit analysis examples

Case studies on Dutch Health Survey and Dutch Labor Force Survey

Mode estimates in sequential mixed-mode design

HS = web, F2F, LFS = web, tel/F2F

	Survey	*Estimate* m_1	*Estimate* m_2
Unemployment rate	LFS	5.6 %	6.7%
% good health	HS	78.0%	75.6%
% smoker	HS	19.9%	29.8%
% obesitas	HS	12.1%	13.9%
% visit to dentist	HS	82.3%	74.5%

Bias intervals for selection and measurement

	Δ_y	*Sign*	*Interval*		*% measurement bias*
			SE	*ME*	
Unemployment rate	1.2%	+	(0.0%, 2.3%)	(-1.1%, 1.2%)	(-92%,100%)
% good health	-2.4%	-	(-4.1%, 0.0%)	(-2.4%,1.7%)	(-100%, 71%)
% smoker	9.9%	+	(0.0%, 4.0%)	(5.9%, 9.9%)	(60%, 100%)
% obesitas	1.8%	+	(0.0%, 3.3%)	(-1.5%, 1.8%)	(-83%, 100%)
% contact dentist	-7.8%	-	(-3.8%, 0.0%)	(-7.8%, -4.0%)	(-100%, 51%)

Extensions: re-interview cost-benefit analysis

Design that is favored under fixed budget and MSE viewpoint (adjusted, m1 only and unadjusted)

RMSE under benchmark m_1

	T	LFS			HS											
		Unemployment ME bias level			Health ME bias level			Smoking ME bias level			Obesitas ME bias level			Dentist ME bias level		
		1.0	.04	-.92	1.0	.14	-.73	1.0	.80	.60	1.0	.09	-.81	1.0	.76	.51
Adjust	3	0.5	0.5	0.5	0.9	0.9	0.9	0.9	0.9	0.9	0.7	0.7	0.7	0.8	0.8	0.8
	7	0.5	0.5	0.5	0.8	0.8	0.8	0.8	0.8	0.8	0.7	0.7	0.7	0.8	0.8	0.8
	19	0.5	0.5	0.5	0.8	0.8	0.8	0.8	0.8	0.8	0.6	0.6	0.6	0.8	0.8	0.8
m1 only	-	<0.1	0.4	0.9	<0.1	1.0	2.0	<0.1	1.0	1.9	<0.1	0.8	1.6	<0.1	0.9	1.8
Unadjust	-	0.5	0.2	0.5	1.7	1.3	1.5	5.0	4.0	3.1	1.3	1.0	1.2	3.9	3.1	2.2

RMSE under benchmark m_2

	T	LFS			HS											
		Unemployment ME bias level			Health ME bias level			Smoking ME bias level			Obesitas ME bias level			Dentist ME bias level		
		1.0	.04	-.92	1.0	.14	-.73	1.0	.80	.60	1.0	.09	-.81	1.0	.76	.51
Adjust	3	0.6	0.6	0.6	0.9	0.9	0.9	0.9	0.9	0.9	0.7	0.7	0.7	0.9	0.9	0.9
	7	0.6	0.6	0.6	0.9	0.9	0.9	0.9	0.9	0.9	0.7	0.7	0.7	0.9	0.9	0.9
	19	0.6	0.6	0.6	0.8	0.8	0.8	0.8	0.8	0.8	0.7	0.7	0.7	0.8	0.8	0.8
m1 only	-	1.1	0.5	0.1	2.4	1.3	0.3	9.9	8.9	7.8	1.8	1.0	0.1	7.8	6.8	5.8
Unadjust	-	0.7	0.2	0.6	1.8	1.3	1.6	5.3	4.3	3.3	1.4	1.0	1.3	4.3	3.3	2.4

Extensions: Latent variable models for measurement effect estimation

Use questionnaire structure or paradata on measurement to estimate measurement bias

- Starting point can be any mixed-mode design, but one mode is chosen as measurement benchmark.
- Psychometric: Questionnaire contains multiple survey items loading on the same latent factors.
- Separate factor(s) may be included for certain answer behaviors such as straightlining or primacy/recency
- Systematic ME components across multiple questions can be identified
- ME on random measurement error can be studied
- Requires adjustment of SE using e.g. covariate-based adjustment methods; risks confounding without strong auxiliary data

Take away messages

- Given different types of auxiliary information, various estimation/adjustment techniques are reported in the literature
- Require choosing benchmark designs, type of adjustment, and cost-benefit criteria
- Typically, measurement biases are removed, while selection biases are preserved
- All estimation/ adjustment techniques have own assumptions that should be thoroughly considered in the scenarios at hand
  - Standard covariate-based estimation using sampling frame or survey data is often invalid
  - Re-interview methodology deals with some shortcomings (e.g. MNAR data)
  - Instrumental variable design too, but makes too strong assumption (representativeness)
  - Time series stabilization easy alternative
- More complex measurement error models using latent variables can reveal additional insights into different types of MEs