IOE, Faculty of Education and Society

Handling Mode Effects in the CLS Cohort Studies

User Guide

November 2024

CENTRE FOR LONGITUDINAL STUDIES

UKRI Economic and Social Research Council

Contact

Questions and feedback about this user guide: clsdata@ucl.ac.uk.

Authors

Liam Wright, George Ploubidis, Richard Silverwood.

How to cite this guide

Wright, L., Ploubidis, G., Silverwood, R. (2024) *Handling Mode Effects in the CLS Cohort Studies*. London: UCL Centre for Longitudinal Studies.

This guide was published in November 2024 by the UCL Centre for Longitudinal Studies.

Data citation and CLS acknowledgement

You should cite the data and also acknowledge CLS following the guidance from cls.ucl.ac.uk/data-access-training/citing-our-data/

Centre for Longitudinal Studies

UCL Centre for Longitudinal Studies (CLS)

UCL Social Research Institute

University College London

20 Bedford Way, London WC1H 0AL

www.cls.ucl.ac.uk

The UCL Centre for Longitudinal Studies (CLS) is an Economic and Social Research Council (ESRC) Resource Centre based at the UCL Social Research Institute, University College London. It is home to a unique series of UK national cohort studies. For more information, visit www.cls.ucl.ac.uk.

This document is available in alternative formats. Please contact the Centre for Longitudinal Studies:

Email: clsdata@ucl.ac.uk

Contents

1 Introduction

Surveys are increasingly moving to mixed mode data collection (Brown & Calderwood, 2020) – for instance, carrying out interviews via face-to-face, telephone, video and/or web. The details of mixed mode data collection differ between studies. In some cases, participants are offered only one survey mode, but the mode offered differs between participants – for instance, as a result of randomization or based on participant characteristics (e.g., predicted likelihood of non-response). In other cases, participants are offered a choice of modes. Modes may also be offered sequentially – for instance, a face-to-face survey may be offered following non-response to other offered modes. Longitudinal studies add a further dynamic element: the set of modes offered – and the rules for offering modes – may differ between survey sweeps.

The potential advantages of mixed mode data collection are lower costs, increased efficiency, and higher participation rates. Participants can be initially offered cheaper modes of data collection (e.g., web survey vs face-to-face interview) and resources can be redirected to recruiting and retaining hard to reach groups. Participants can also complete surveys in modes that are most convenient or suitable, given their situations – for example, some older adults may be unable to complete web surveys and working parents may not have the time to complete a face-to-face interview.

These advantages do not come without drawbacks, however. Specifically, responses may differ systematically between survey modes due to differences in how items are measured, including the context in which they are presented. For instance, the presentation of a survey item either orally or visually may influence responses, sensitive information may be reported more accurately when provided anonymously (e.g., by web survey compared with face-to-face interview) and complex information may be reported more accurately when an interviewer is present. Differences in responses arising from differences in measurement are termed ***mode effects***. Unaccounted for, mode effects can be a problem for obtaining accurate and unbiased estimates, both for descriptive and inferential statistics. For instance, changes in survey mode between sweeps may bias estimates of change (Cernat & Sakshaug, 2021) and mode effects can induce associations between variables (e.g., the correlation between mental health and employment status where measurements of both are influenced by survey mode; Buelens & Brakel, 2017; Goodman et al., 2022).

Researchers may be tempted to try to account for mode effects by simply adding an indicator variable for mode into analyses of mixed mode data. However, because participants are not randomly distributed across modes – even where randomized to mode *offered*, individuals choose whether they participate or not – observed differences in responses between modes are a combination of mode effects (*how* an item is being measured) and selection effects (*who* is being measured). Selection effects can confound mode effects. Adding a mode indicator variable into analyses does not address this. In fact, in some situations, it can increase bias.

Whether responses exhibit mode effects depends on the specific survey item – easily recalled, objective characteristics, such as age, may be reported accurately regardless of mode used, while subjective and socially sensitive characteristics, like mental health, may not be. Further, whether mode effects lead to biased results depends on the particular analysis being carried out – the variables being used, the relationships between them, and their relation to mode selection. For example, a regression of mental health on sex may yield an unbiased beta coefficient if sex is not predictive of mode of survey completion or subject to mode effects (though standard errors will be larger than they otherwise would be). Given this, it is not possible to give concise and definitive recommendations for handling mode effects that apply in all situations. Instead, the aim of this document is to introduce frameworks for thinking about mode effects in your own analyses of Centre for Longitudinal Studies' (CLS) cohort data.

This document can be read as standalone sections, and there is signposting throughout; please do not be deterred by its overall length. The outline of this document is as follows. In Section 2, we enumerate the elements of mixed mode data collection that have appeared in CLS's cohorts. In Section 0, we discuss characteristics relevant to predicting *a priori* whether a specific survey item will exhibit mode effects. In Section 4, we introduce frameworks for thinking about the consequences of mode effects in a variety of settings. In Section 5, we provide empirical evidence on (a) mode effects in Sweep 9 of the National Child Development Study (NCDS), which embedded an experimental mixed mode web and telephone survey design, (b) mode selection effects in Sweep 9 of NCDS and Sweeps 4-8 of Next Steps, and (c) a review of the empirical literature on mode effects from four major social science and health surveys. In Sections 6 and 7, we provide recommendations for accounting for mode effects. Finally, in Section 8, we show two worked examples

of adjusting for mode effects using data from Sweep 9 (55y) of the NCDS and Sweep 6 (18/19y) of Next Steps.

2 Mixed Mode Elements of CLS' Cohort Data Collections

Several of the studies run by CLS have included elements of mixed mode data collection, both between individuals within a sweep and within individuals across sweeps, as well as between individuals *across studies* – important for conducting cross-cohort analyses of CLS data. Below, we outline the main elements of within and between sweep mixed mode data collection.

2.1 Within-Sweep Mixed Mode Data Collection

Within sweeps, mixed mode data collections include: [1]

1. Millennium Cohort Study (MCS): Sweep 6 Time Use Diaries (Age 14y)
    - Cohort members were offered the choice of web and mobile app data collection modes, with a paper version also available if the cohort member was unable to use either of these modes.
2. 1970 British Cohort Study (BCS70): Sweep 3 Cognitive Ability Tests (Age 16y)
    - Data collection at the age 16y sweep was conducted during a teacher strike. Consequently, some participants completed cognitive tests at home rather than at school and almost half of participants did not complete the tests at all.
3. Next Steps: Sweeps 5-9 Interviews (Ages 17/18y – 32y)
    - Sequential mixed mode approaches were used in each sweep. In Sweeps 5-8, a web survey was offered initially, followed by telephone and, finally, face-to-face interview. In Sweep 9, a web survey was offered initially, with non-respondents then able to choose from face-to-face, telephone, or video interview or web survey. A shortened web survey was then offered to remaining non-respondents.
4. 1958 National Child Development Study (NCDS): Sweep 9 Interview (Age 55y)
    - A sequential mixed mode approach was used primarily (web initially offered followed by telephone interview), but the survey also embedded

[1] There are further nuances to each of these, such as some individuals automatically being offered face-to-face interview. Readers should refer to study user guides for more detail.

a mode experiment, with 1,499 of 10,558 cohort members randomly allocated to telephone-only data collection (Goodman et al., 2022).

5. CLS COVID-19 Surveys: Sweep 3 (February – March 2021)
    - During the COVID-19 pandemic, participants of the MCS (parents and cohort members), BCS70, NCDS and Next Steps were invited to complete surveys on their pandemic experiences. The third sweep was conducted using a sequential mixed mode approach with a subset of participants offered web initially followed by telephone interview. Other participants were allocated to web interview only.

The proportions of participants in each mode for (1)-(5) are displayed in Figure 2.1.[2] Note, most sweeps also include an element of mixed mode within a sweep *within an individual* – for instance, combining face-to-face interview with self-completion modules for particularly sensitive questions. Though, in this case, different sets of questions are typically asked in each mode. For instance, Sweep 7 of the MCS contained a web survey of cohort members following their face-to-face interview.

Note, web surveys add additional complexity as participants may be able to respond on different types of device (e.g., phone, tablet, or computer). The device used can affect how questions are presented or where the survey is completed and thus could influence how individuals respond (so called 'device effects'). Please refer to study user guides on the devices web participants were instructed or explicitly limited to respond using. In some cases, variables are available with the data capturing device type (e.g., `N9DEVICE` at Sweep 9 of the NCDS). Advice in this user guide applies to device effects as well as mode effects.

2.2 Between-Sweep Mixed Mode Data Collection

Modes have also varied across survey sweeps within a study. Between-sweep mixed mode data collection includes:

1. MCS
    - In Sweeps 1-6, parents completed interviews via face-to-face interview. In Sweep 7, web survey was used, though interviewers encouraged

[2] Only 9 participants in Sweep 9 of Next Steps participated via video interview. We do not include this figure in the plot for brevity.

parents to complete this while the interviewers were in the house interviewing cohort members.

- Cohort members completed self-complete questionnaires by paper in Sweeps 4-5 and (confidentially) using the interviewer's tablet in Sweeps 6-7. Audio (Sweeps 5) or interviewer (Sweeps 5-7) was additionally used where necessary (e.g., for cohort member with literacy issues). In Sweep 7 a follow-up self-complete web survey was administered within ten days of the main home interview, which could be completed on any digital device.
    i. In Sweeps 2-3, older siblings completed a self-completion survey on paper. Combining responses from cohort members and their older siblings may mix modes.

2. Next Steps
    - Sweeps 1-4 (13/14y – 16/17y) were carried out by face-to-face interview. Sequential mixed mode approaches were used in Sweep 5-9 (17/18y-32y) with participants able to respond via web survey, telephone or face-to-face interview and, latterly, video interview (age 32y).
3. BCS70
    - Sweeps 1-4 (0y-16y), 6-7 (30y-34y) and 9-10 (42y-46y) were carried out by face-to-face interview. Sweep 5 (26y) was carried out by postal paper questionnaire and Sweep 8 (38y) was carried out by telephone.
4. NCDS
    - The Sweeps 0-6 (0y-42y), the Biomedical sweep, and Sweep 8 (50y) were carried out by face-to-face interview. Sweep 7 (46y) was carried out by telephone and Sweep 9 (55y) was carried out by telephone or web (see above).
5. CLS COVID-19 Surveys
    - The initial two sweeps were carried out by web interview, while Sweep 3 was carried out by telephone or web interview. The use of web or telephone interview differs from pre-COVID-19 sweeps in each of the surveys, except Sweeps 5-8 (17/18y – 25y) of Next Steps, Sweep 8 (38y) of the BCS70, and Sweeps 7 (46y) and 9 (55y) of the NCDS.

The series of modes participants used in Sweeps 1-8 of Next Steps are displayed in Figure 2.2.

2.3 Planned or Ongoing Data Collections using Mixed Mode Designs

Future and ongoing data collections including mixed mode elements are also planned:

- MCS Sweep 8 Interview (Age 23y): Web and face-to-face.
- BCS70 Sweep 10 Interview (Age 51y): Web, face-to-face, and video interview.
- NCDS Sweep 10 Interview (Age 62y): Web, face-to-face, and video interview.

In each case, the use of mixed modes for the data collections will implying differences in mode between sweeps.

Note, where there is no overlap in the modes used between sweeps (e.g., telephone or face-to-face was used at sweep *x*, but web survey was used at sweep *x+1*), there may be no information contained in the data itself that can be used to correct the mode effects as the counterfactual cannot be directly estimated (i.e., there are no web survey responses at sweep *x*). However, external information could be used to make informed guesses and simulate data (see Section 6).[3]

[3] Responses regarding time-invariant information (e.g., age at entry to country for migrants, etc.) could potentially be used to estimate mode effects between sweeps. However, these would be confounded by recall bias as responses from different time points would be used. Further, items on time-invariant information may be limited to purely factual information and so would be expected to exhibit smaller – or non-existent – mode effects (see Section 0).

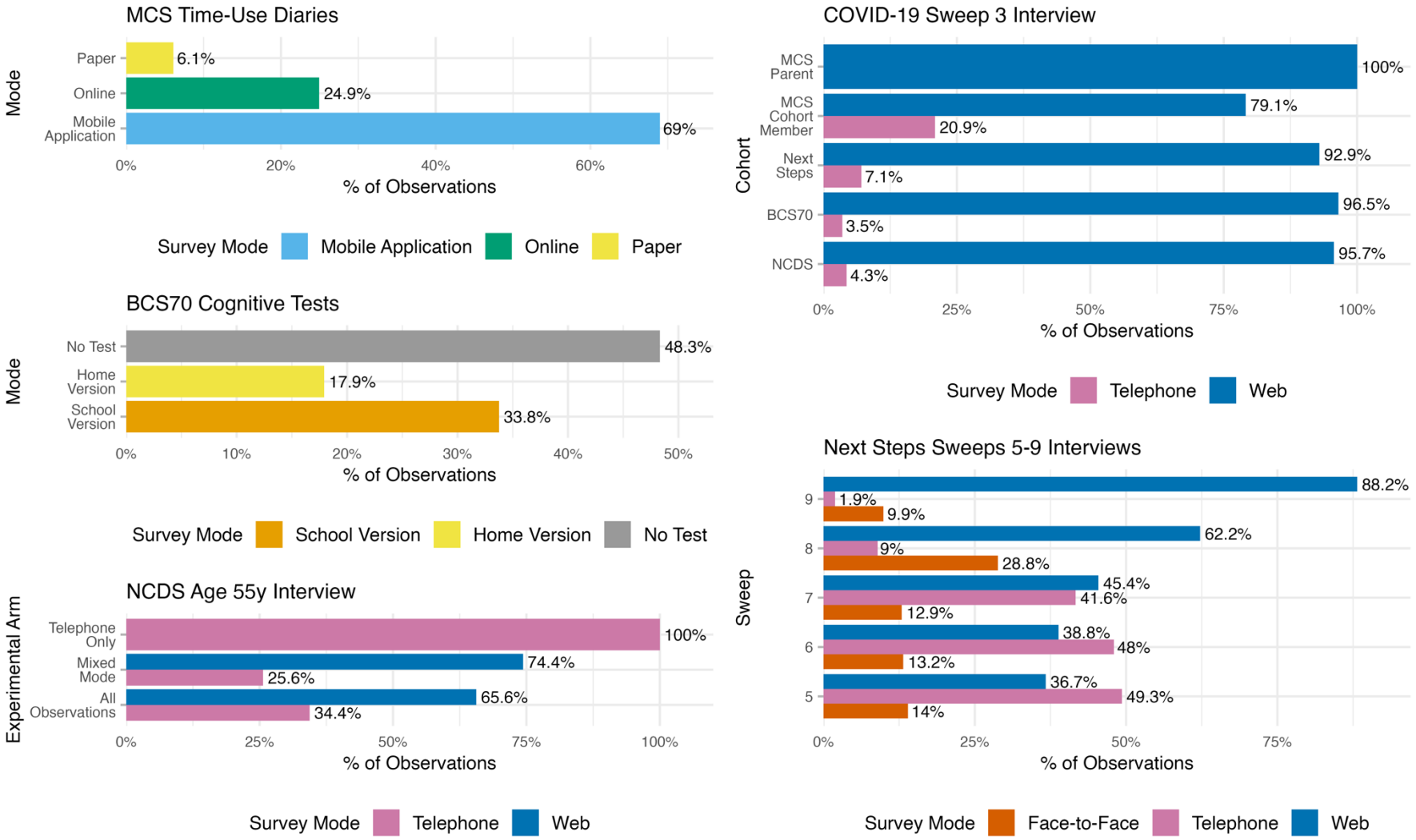

Figure 2.1: Proportions of respondents using each mode for the mixed mode elements of CLS' cohort studies

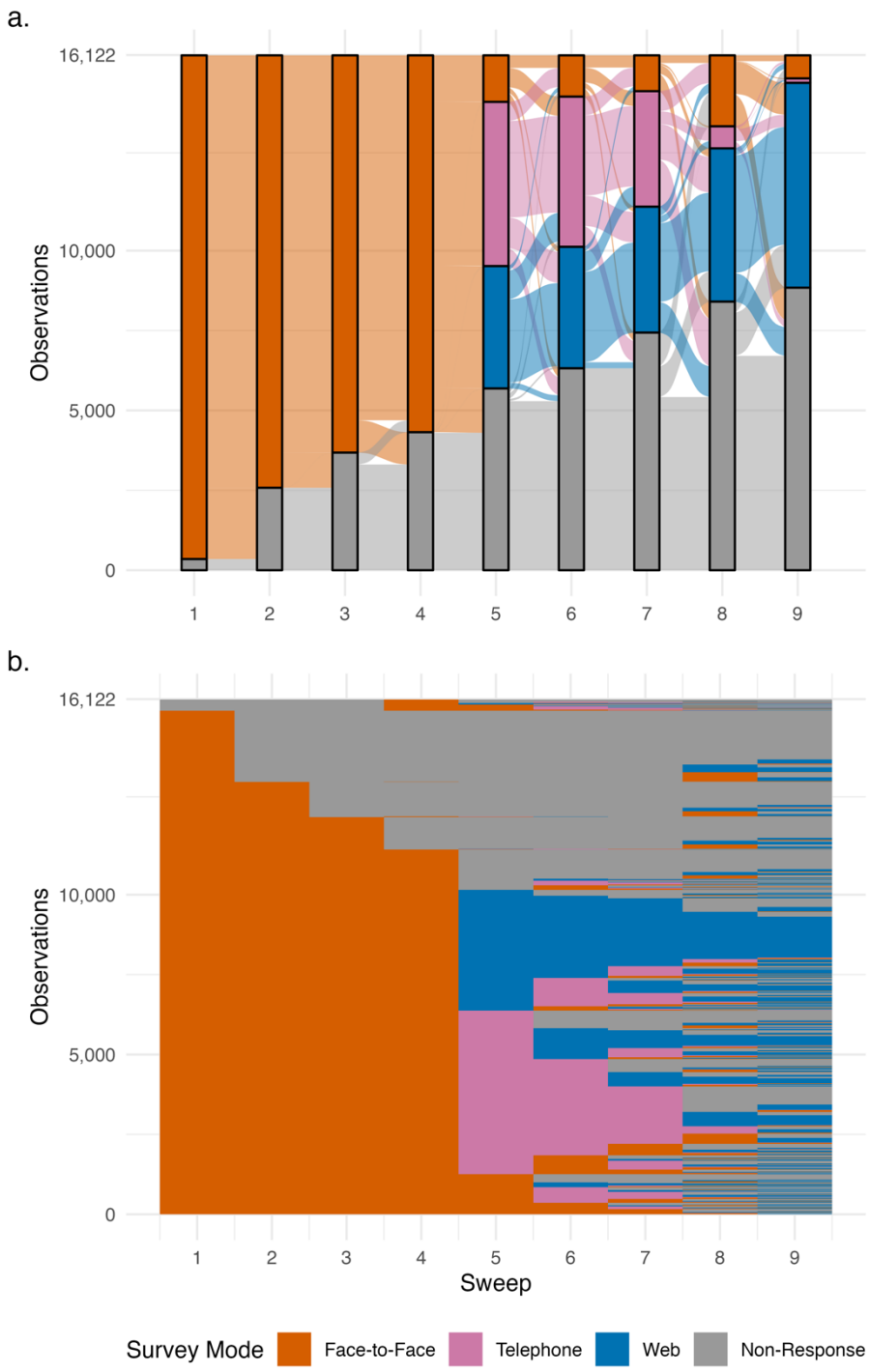

Figure 2.2: Movement between modes in Sweeps 1-9 of Next Steps. (a) Alluvial diagram of mode switching between adjacent sweeps. (b) Sequence plot of series of modes used by Next Steps participants. In (a), sets of participants sharing the same combination of modes in adjacent sweeps (e.g., Sweeps 4 and 5) are represented by a coloured band. The width of this band is proportional to the number of participants in the set and the colour represents the mode used in the earlier sweep, while the colour of the box the band feeds into represents the mode used in the later sweep. For instance, the blue band that goes from the blue to pink boxes in Sweeps 4 to 5 represents the group of participants who responded by web mode in Sweep 4 and

telephone mode in Sweep 5. In (b), each participant is represented by a colour row of data and the colour at each sweep (x-axis) represents the mode used at that sweep. For instance, row that are orange in Sweep 1-4 and switch to blue at Sweeps 5-9 represents individuals who participated by face-to-face interview in the first four sweeps and web questionnaire in the latter four sweeps. Note, in Sweep 9 nine cohort members participated via video interview. To avoid clutter, we do not show these on the plots.

3 A Framework for Predicting Which Survey Items are Liable to Mode Effects

Differences in responses between survey mode can be influenced by mode effects (i.e., measurement error; differences in response induced by overall presentation of items) or mode selection (i.e., differences in the *people* who answer in each mode). As Section 4 will show, accounting for survey mode is only necessary where differences in responses between mode are influenced by mode effects, rather than selection effects alone. In this section, we describe the characteristics of survey items that are liable to mode effects and discuss the potential consequences of these characteristics for the *distribution* of responses. We also briefly discuss the consequences of mode for item and survey non-response.

3.1 Characteristics of Survey Items Susceptible to Mode Effects

Based on their review of the mode effects literature, d'Ardenne and colleagues (2017) created a framework for predicting the likelihood a survey item will exhibit mode effects, *a priori*. They note three overlapping sets of factors that increase the risk of mode effects: interviewer effects; satisficing; and question and answer presentation issues.

Interviewer effects refer to differences in responses due to an interviewer being present or not. Risk factors for interviewer effects are fear of disclosure, socially desirable reporting, and positivity bias. A respondent may be less likely to report embarrassing, illegal or illicit behaviour or negative opinions in the presence of an interviewer to present themselves more favourably or out of concern the interviewer might share their responses with others.

Interviewer effects can come in degrees – people may feel less comfortable revealing sensitive information in a telephone interview than a web survey, where there is no interviewer at all, but more comfortable than in a face-to-face interview, where an interviewer is physically present. Interviewer effects are the reason why surveys typically use self-completion modules to elicit sensitive information, though it should

be noted that physical presence alone can be sufficient to influence responses (Burkill et al., 2016). In telephone surveys, technical limitations mean responses may need to be given aloud, even for sensitive questions, breaking anonymity. In this case, response categories are sometimes amended (e.g., to 'yes' or 'no' answers) so that privacy is otherwise maintained.

Satisficing refers to the tendency for respondents to give answers that are "good enough", rather than expending additional cognitive, or other, effort to understand and consider the question in full and provide the most accurate response possible. Interviewers can explain complex tasks, provide additional information, and increase motivation to perform tasks thoroughly. Additional information may not be available or remain unused in other modes, such as in self-completion modules and web surveys.

The risk of satisficing is assumed to be greater for complex or difficult questions than straightforward or easy questions (d'Ardenne et al., 2017). d'Ardenne et al. (2017) list five factors that can influence the risk of satisficing: complex question stems, provision of extra information, computations, document consultation, and open questions. Evidence that survey respondents do not always maximise effort or provide fully considered answers is abundant (Roberts et al., 2019). For example, in Wave 7 of the Understanding Society Innovation Panel, more than 60% of participants answered a question on overall life satisfaction within ten seconds.[4]

Question and answer presentation refers to how the respondent receives information and reports their answer. For questions with multiple response categories (e.g., Likert scales), in oral presentation, the participant may be more likely to report the categories that come toward the end as these are heard last. In visual presentation, the same participant may be more likely to report categories that come towards the start as these are read first and the respondent may satisfice and not read the full question. These are known as recency and primacy effects, respectively (Krosnick & Alwin, 1987). Other salient differences include the possibility the respondent has to scroll to read the full question in a web survey (which they may not opt to do nor realise they have to do), repetition of response scales (which may provoke individuals to

[4] Own calculation. Unweighted.

provide the same response to each ['straightlining']; Kim et al., 2019), and use of scales with mid-points (which draw the eye in visual modes).

It is possible that individual survey items feature multiple characteristics susceptible to mode effects simultaneously. Box 3.1 displays three items from the CASP quality of life measure (Hyde et al., 2003), which was given to telephone and web respondents at the Sweep 9 (age 55y) interview of the NCDS (Wiggins et al., 2017). In the web survey, these items were displayed on screen in a grid by row, with checkboxes in columns and the response options listed as column headers. In the telephone interview, items were presented (and answered) orally with the response categories repeated after each question.

Box 3.1: Three items from the 6-item version of the CASP quality of life measure included in the age 55y NCDS web survey and telephone interview.

Here is a list of statements that people have used to describe their lives or how they feel. We would like to know how often, if at all, you think each applies to you? Please say whether each applies to you often, sometimes, not often or never.

1. I feel full of energy these days
2. I feel that life is full of opportunities
3. I feel that the future looks good for me

Given the topic sensitivity, telephone responses to these items are likely to be subject to interviewer effects. This should bias responses towards overstating wellbeing levels relative to the web survey where an interviewer was not present. However, as the telephone interview was presented orally, responses could alternatively be biased towards response categories heard last (a recency effect) understating wellbeing relative to the web survey in which responses may be biased towards those read first (a primacy effect).[5] Which process is more important is an empirical question. Thus, for items like these, an awareness of the evidence on the overall direction of bias is important to usefully predict mode effects, *a priori*. (In practice, respondents in the

[5] Web responses could moreover be influenced by satisficing if participants rush through and give the same answer to each item. The item is likely to be subject to satisficing in general, too. Individuals may use an availability heuristic, prioritising more recent experience or other information that spring immediately to mind (Kahneman, 2012; though see Yap et al., 2017). Whether this would lead to understating or overstating wellbeing is unclear.

telephone interview reported higher wellbeing, on average; Goodman et al., 2022; also see Section 5.)

3.2 Mode Effects on the Distribution of Responses

Mode effects can entail consequences for the distribution of responses, rather than simply mean differences between survey modes (Clarke & Bao, 2022). For Likert items, responses may exhibit different skewness in one mode compared with another – for instance, when presented orally, responses may be clustered at categories stated last (Krosnick & Alwin, 1987). Mode effects like these generate differences in multiple moments and parameters of the response distribution (e.g., mean, variance and kurtosis, median and mode), and could have consequences for statistical analysis, such as biased estimates and reduced statistical power.

The size of the mode effect could also differ between individuals. Again, this can have consequences for the distribution of responses. Importantly, mode effects may differ according to participant characteristics that are related to variables of substantive interest. For instance, the consequences of satisficing may be greater among those who struggle processing complex information and interviewer effects may be weaker among those who are more disagreeable. This may be particularly important for studies on cognitive ability or personality traits. What is considered socially desirable may differ between individuals, for example by sex (Burkill et al., 2016) or by generation, the latter having implications for cross-cohort analyses, in particular.

Mode effects could also differ according to the underlying level of the characteristic the survey item is intended to capture. This is especially true for censored or non-continuous variables. For binary variables, individuals can only be misclassified in one direction (i.e., false positive or false negative). For Likert items, individuals whose underlying value is at an extreme can only be misclassified in the opposite direction to that extreme, while those with intermediate underlying values can be misclassified in either direction. This can also have consequences for accurately accounting for mode effects, as discussed further in Section 4.

3.3 The Impact of Mode on Item and Survey Non-Response

Interviewer effects, satisficing, and question and answer presentation may not only influence responses given but also item-response rates. For instance, participants

may be less willing to discuss sensitive topics in the presence of an interviewer and may thus refuse sections of the survey. Participants may also have less motivation to complete tests in an anonymous mode, such as a web survey. Item non-response may also be related to the underlying value of the characteristic being elicited; a person who would struggle with a cognitive test may choose to avoid one if they can. In other words, mode effects may mean that data are missing not at random. Effects of mode on item non-response can be large. In their analysis of the mixed mode experiment from Sweep 9 (55y) of the NCDS, Goodman et al. (2022) observe differences in item-response between web and telephone surveys as high as 9.7 percentage points (pp.; employer provided pension type; 95% CI = 7.0 pp., 12.5 pp.).

Effects of mode on unit non-response can also be sizeable. The mixed mode group at age 55y in the NCDS had a 5 pp. higher response rate (82.8% vs 77.8%) than the telephone only group. It is worth noting that effects of mode on unit non-response can reflect factors other than respondent decision making (e.g., refusing to participate when the mode offered is unappealing), such as administrative issues like correctly identifying and following potential participants. Response rates will be lower if relevant contact details are outdated or incorrect (e.g., address details for face-to-face interview or email for web survey).

3.4 An Application to Survey Items from the NCDS and Next Steps

In this section, we apply d'Ardenne et al.'s (2017) framework to two survey items from the NCDS and Next Steps. Box 3.2 displays an item on self-rated general health from Sweep 7 (age 46y) of the NCDS, a telephone-only survey, unlike previous sweeps in which a similar question was asked via face-to-face interview.

Box 3.2: Self-rated health survey item from Sweep 7 (age 46y) of the NCDS

Please think back over the last 12 months about how your health has been. Compared to people of your own age, would you say that your health has on the whole been...

1. …excellent
2. Good
3. Fair

4. Poor or
5. Very poor?

Question and answer presentation in the telephone and face-to-face interviews is similar in that the item is stated orally; while this may bias towards later response categories, it should be similarly valent in both modes. Given an interviewer is not physically present in the telephone interview, responses may be subject to stronger interviewer effects in the face-to-face interview, biasing towards reporting better health in that mode. Though, as the participant is directly observed in the face-to-face interview, this impulse may be tempered by a motivation to give a plausible answer. Additionally, answering the question accurately involves considering health over a relatively long timespan (twelve months), requiring some cognitive effort. The greater opportunity physical presence affords for developing a positive relationship with an interviewer may motivate the participant to expend effort to answer the question accurately. As with the CASP-6 items shown in Box 3.1, while we may expect a mode effect – which in this case would bias estimates of change between NCDS survey sweeps – it is difficult to predict the direction, *a priori*. Instead, existing empirical evidence must be used where available.

Box 3.3 shows an item on whether the participant has any children from Sweep 8 of Next Steps, which used a sequential mixed mode design (web followed by telephone then face-to-face interview). This item is unlikely to exhibit mode effects as it elicits factual information that should be easy to recall and salient in participants' lives. Moreover, potential ambiguity in the question has been minimized by the interviewer providing a detailed definition of children. The response categories ('yes' and 'no') are also definite and even allow for some errors in calculation (e.g., a person could still answer 'yes' if they misremember their exact number of children).

Box 3.3: Question on parenthood from Sweep 8 (age 25y) of Next Steps

Do you have any children?

Please include any adopted children, step-children or foster children of whom you consider yourself to be a parent in addition to your own biological children. Please also include children who do not currently live with you.

1. Yes

2. No

While it is possible to come up with instances in which a participant would want to misreport their number of children (e.g., where the participant is parent of a child which the people they live with do not know about), these are likely to be rare and unlikely to make an important difference at an aggregate level. Identifying items that are unlikely to exhibit mode effects is useful as they can potentially be used as negative control outcomes (Lawlor et al., 2016; Lipsitch et al., 2010) to establish whether it is possible to unbiasedly estimate mode effects in a given survey – any association between the item and survey mode should reflect mode selection (see Section 8).

4 A Framework for Understanding the Consequences of Mode Effects

To describe the consequences of mode effects, in what follows, we use a counterfactual approach based upon the Potential Outcomes Framework (Morgan & Winship, 2015; Rubin, 2005). In this framework, a mode effect is the difference between the observed measurement and the measurement that would have been observed had the participant answered the survey in another specific mode. The latter is counterfactual because it is counter-to-the-facts and is necessarily unobserved. We use Directed Acyclic Graphs (DAGs) to visually represent the consequences of mode effects. DAGs are an exceptionally helpful tool for mapping causal relationships and understanding potential sources of bias in data. The next section provides an introduction to DAGs for readers who are unfamiliar with them.

In discussing the consequences of mode effects, we assume that the aim is to obtain consistent and unbiased estimates of what would have occurred if a single, specific survey mode was used for all cohort members or across all survey sweeps. We discuss three types of analysis:

1. Descriptive statistics (e.g., means, proportions, and standard deviations for single variables or bivariate correlations between pairs of variables).
2. Associations and causal effects (e.g., estimates from regression models on relationships between variables, potentially controlling for confounders).
3. Longitudinal analyses (e.g., trajectory modelling and estimates of within-person change in a variable).

This list is not exhaustive. The following is intended as a framework for thinking about mode effects in the context of your own analysis of CLS data.

4.1 An Introduction to Causal Directed Acyclic Graphs (DAGs)

DAGs visually represent causal relationships. In DAGs, variables are represented by nodes and causal relationships by directed arrows. A graph is a DAG if it fulfils two criteria:

1. It is acyclic – it is not possible to follow arrows forward and arrive at a variable already reached.
2. It is complete – the common causes (observed or unobserved) of any two variables in the DAG also appear in the DAG.

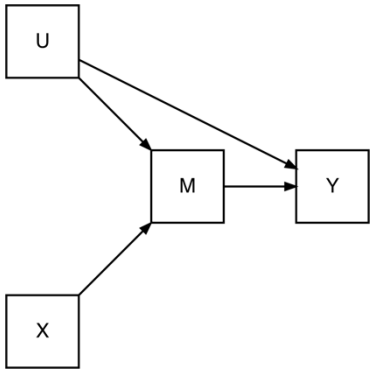

Figure 4.1: An Example DAG

Figure 4.1 shows an example DAG. It contains four separate variables – X, U, M and Y – and four (directed) arrows. The variables are connected by *paths* comprising one or more adjoining arrows. Paths are *direct* (passing through no intermediate variables; e.g., U $\rightarrow$ Y) or *indirect* (passing through one or more intermediate variables; e.g., X $\rightarrow$ M $\rightarrow$ Y) and do not necessarily flow in the direction of causality; for instance, M $\leftarrow$ U $\rightarrow$ Y represents one path between M and Y.

An essential characteristic of DAGs is that they follow a set of straightforward rules. These rules imply statistical relationships. Paths are either 'open' or 'closed'. Variables that are connected by an 'open' path should be associated, while variables that are connected only by 'closed' paths should not. Open paths between two variables are not necessarily causal, in the sense that they are 'directed' with arrows pointing in the same direction. Whether a path is open or closed depends on the variables being conditioned upon.

Figure 4.1 contains six directed paths representing causal association: X $\rightarrow$ M, X $\rightarrow$ M $\rightarrow$ Y, M $\rightarrow$ Y, U $\rightarrow$ M, U $\rightarrow$ M $\rightarrow$ Y, and U $\rightarrow$ Y. These paths are open but can be closed by conditioning upon intermediate variables (*mediators*), where they exist – i.e., the paths X $\rightarrow$ M $\rightarrow$ Y and U $\rightarrow$ M $\rightarrow$ Y can be closed by conditioning upon M. The path from M to Y via U (M $\leftarrow$ U $\rightarrow$ Y) is a non-causal 'backdoor' path. Such paths are said to be *backdoor* as they exit the back of the initial variable. The path M $\leftarrow$ U $\rightarrow$ Y

is open but can be closed by conditioning upon U. This path represents the belief that U is a confounder for the relationship between M and Y.

Finally, Figure 4.1 contains a closed path from X to Y via M and U ($X \rightarrow M \leftarrow U \rightarrow Y$). M is a descendent (a consequence) of both X and U and is known as a 'collider' (the paths from X and U *collide* at M). Colliders close paths unless they (or a descendent of the collider) are conditioned upon. An open collider path generates 'collider bias', a form of selection bias. Examples of collider bias abound and even apply in Randomized Controlled Trials (RCTs). Take Hollywood actors (Elwert & Winship, 2014): physical attractiveness and acting ability may be uncorrelated in the general population, but both increase the chances of becoming a star. Among professional actors, the less than prepossessing must have a surfeit of skill, otherwise they wouldn't have been hired. Those lacking in talent must have a face to make up for it. Hence, attractiveness and acting ability are negatively correlated once being a professional actor is conditioned upon.

Despite its simplicity then, we can read a lot of information off the DAG in Figure 4.1, so long as we accept the assumptions embedded within it. The association between M and Y will be confounded in observational data, but we can deal with this confounding by conditioning upon U. X only has a causal effect on Y via M; X is therefore a source of exogenous variation in M and could be used as an instrumental variable for M – particularly helpful if data for U are not available. This is the power of DAGs.

For more information on DAGs, see one of the many tutorials on that are now available (Digitale et al., 2022; Hernan & Robins, 2023; McElreath, 2020; Pearl & Mackenzie, 2018; Williams et al., 2018). We particularly recommend Miguel Hernán's (2018) pellucid free online course, *Causal Diagrams: Draw Your Assumptions Before Your Conclusions*.

4.2 Representing Mode Effects using DAGs

In our DAGs, we represent mode effects as a form of systematic measurement error (VanderWeele & Hernan, 2012). There is a variable, Y, that represents the true value of the construct we wish to measure. Y is latent and is not observed directly but instead captured by survey item(s) – for instance, Y may be a dimension of mental health that we have tried to assess using a battery of questions. The observed value, Y*, is a

function of this latent variable and the survey mode, M. Figure 4.2 shows the basic set-up. To simplify the notation, we assume there are two survey modes (i.e., M is a binary indicator variable) and that measurement error is not present in the reference survey mode. That is, Y* = Y for those in the reference mode, but Y* ≠ Y for those in the alternate mode. Though there may in fact be measurement error under both modes, in practice researchers are likely to be particularly interested in the counterfactual: 'What would responses have been if only survey mode *M* was used?'.

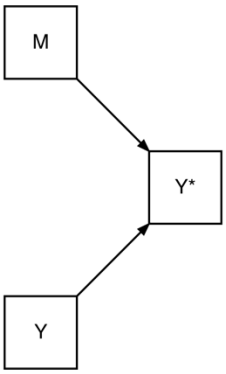

Figure 4.2: Directed Acyclic Graph (DAG) of Mode Effect

4.3 Mode Effects Only

Figure 4.2 above represents a relatively straightforward situation where there is a mode effect, but selection into mode is unrelated to Y and the size of the mode effect does not vary according to the value of Y (in other words, mode is independent of Y). This would be the case if the mode effect is a source of random measurement error or is constant across all individuals and if the factors influencing selection into mode (e.g., chance, where the mode offered is randomised and does not influence participation rates) are not causes or consequences of Y.

The consequences of this situation depend on the analysis to be performed and the form of the mode effect. Without correction, in univariate analyses of Y* (e.g., descriptive statistics), the mean and variance will be biased relative to univariate analyses of Y, if we could perform them. A straightforward solution in this situation is to calculate these statistics using only observations from the reference survey mode. However, this option does not make full use of the data and can be inefficient relative to methods which estimate the counterfactual for observations from the alternate

mode; this inefficiency can be important if few participants are in the reference survey mode.

Where mode effects are fixed (i.e., they are the same for every individual), another method is to estimate the mode effect by regressing observed Y* upon mode M (an indicator variable: 0 = reference mode, 1 = alternate) and then subtract this estimate from observed Y* for those in the alternate mode to obtain an estimate of the counterfactual, Y, for these people.[6] These values can then be used to calculate descriptive statistics or in other forms of analysis. Note, as the counterfactual is estimated, uncertainty in the estimate should be propagated into the final statistics (e.g., by producing confidence intervals with bootstrapping). This is discussed further in Section 6.

This procedure will produce biased estimates of variance and associated statistics (e.g., standard deviation, IQR, and centile values) if the mode effect is heterogeneous (i.e., individuals differ in the extent to which mode influences their answers) as it only accounts for mode effects upon the mean and not on variability. In this case, there is a trade-off between bias and efficiency – using more observations from the alternate mode increases precision but biases estimated variance in the (counterfactual) reference mode. The trade-off depends on the level of heterogeneity in the mode effects and the size and relative proportion of the sample in each mode. More complex procedures are possible (e.g., generalized additive models for locations, scale, and shape [GAMLSS]); these are discussed further in Section 6.

As the variance of Y* will not be equal to the variance of Y, correlations between Y* and variables not related to selection into mode or subject to mode effects will be biased (e.g., variables X and Z in Figure 4.3a-b). Beta coefficients from regression models of Y* upon X (a cause of Y) will be unbiased as the potentially biasing path M $\rightarrow$ Y* $\leftarrow$ Y $\leftarrow$ X is closed because Y* is a collider (see also data simulations in Figure 4.4a-b). However, statistical power may be lower relative to a situation in which a single mode was used. The intuition behind the lack of bias is that, because X is uncorrelated with M, each unit increase in X has the same effect on Y* as it does on Y.

[6] This procedure will not reduce inefficiency in estimates of the mean but reduces inefficiency in estimates of variance.

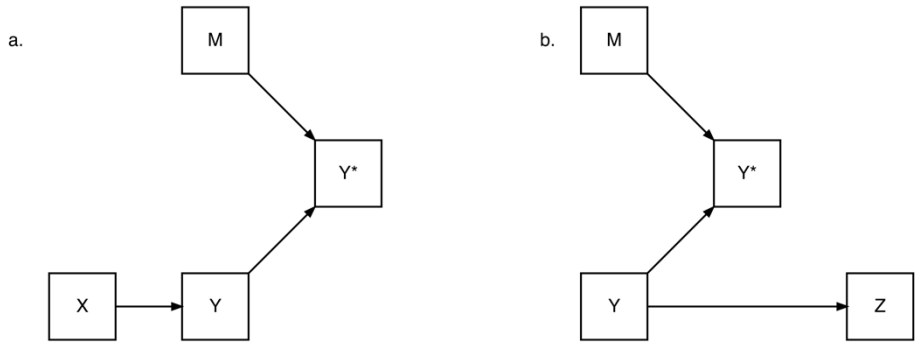

Figure 4.3: Directed Acyclic Graph (DAG) of a simple mode effect where the size of the mode effect is unrelated to a variable of interest, Y, and Y is not related to selection into survey mode.

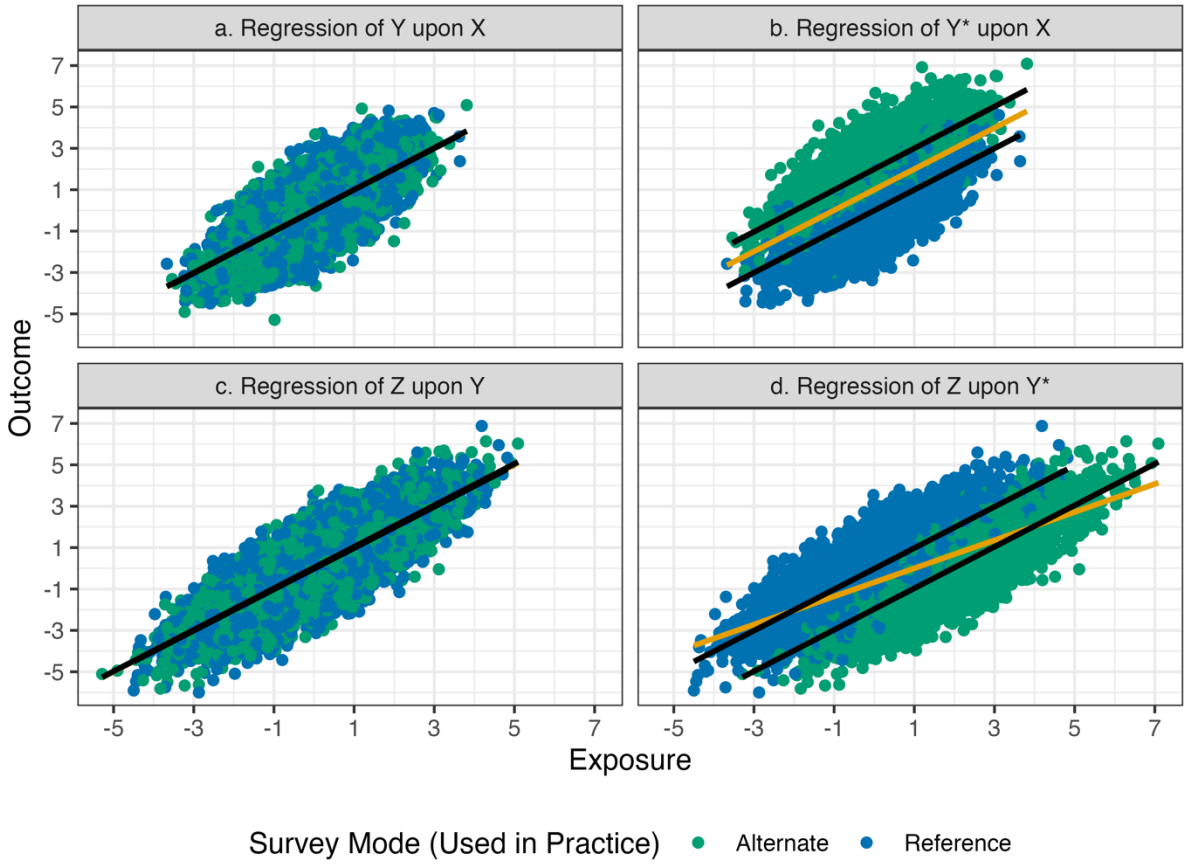

Figure 4.4: Associations between variables X, Y, and Z as calculated via linear regression and reflected in the relationships in Figure 4.3. Y is subject to a (fixed) mode effect. Orange lines represent regression lines estimated from the total sample. Black lines indicate regression lines when stratifying by survey mode. The colour of the points reflects the mode the survey participant used in actuality. Y* is the observed value of Y, while Y is the (possibly, counterfactual) value that would have been observed had the survey participant used the reference mode. Panel (a) shows the results of a regression of (counterfactual) Y upon X. Panel (b) shows the results of a regression of (observed) Y* upon X. Data simulations for panel (a) and (b) were

generated to reflect relationships in Figure 4.3a. Panel (c) shows the results of a regression of Z upon (possibly counterfactual) Y. Panel (d) shows the results of a regression of Z upon (observed) Y*. Comparing panels (a) and (b) there is no bias in the regression estimate when using Y* as an outcome variable. Comparing panels (c) and (d), the regression estimate is attenuated to the null when using Y* as an exposure variable. However, an unbiased estimate is obtained stratifying by survey mode. Data simulations for panel (a) and (b) were generated to reflect relationships in Figure 4.3b.

Beta coefficients from regression models of Z (a consequence of Y) upon Y* (Figure 4.3b) will be biased towards the null as conditioning upon Y* opens the path M $\rightarrow$ Y* $\leftarrow$ Y $\rightarrow$ Z. Intuitively, there is a source of variation in Y* that does not impart a causal effect upon Z. Each unit increase in Y* will therefore be more weakly associated with Z than Y is (see data simulations in Figure 4.4c-d). This bias is known as regression dilution or attenuation bias (Hutcheon et al., 2010). The bias can be removed by stratifying analyses by survey mode or adding mode as a control variable.

Attenuation bias also creates issues where variables exhibiting mode effects are used as control variables. Figure 4.5 shows a situation in which Y is a confounder for the association between X and Z. Controlling for Y* again opens the path M $\rightarrow$ Y* $\leftarrow$ Y $\rightarrow$ Z. Survey mode causes variation in Y* that does not reflect the underlying value of Y which means that some residual confounding remains (i.e., two individuals with the same Y* may differ on their underlying Y). The degree of residual confounding increases with the size of the mode effect.

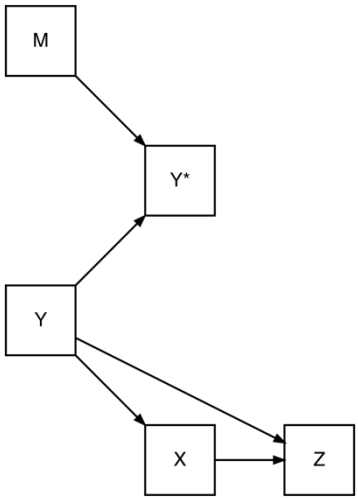

Figure 4.5: Directed Acyclic Graph (DAG) of a situation involving mode effects on a variable Y that confounds the association between X and Z.

4.4 Mode Effects with Mode Selection

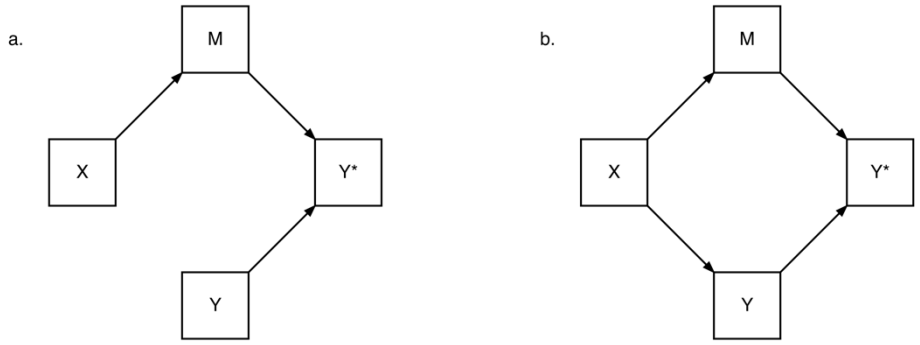

Figure 4.6: Directed Acyclic Graph (DAG) of a situation involving (i) mode effects on a variable Y and (ii) mode selection according to a variable X. In panel (a), variable X has no causal effect on variable Y. In panel (b), variable X has a causal effect on variable Y.

Figure 4.6a represents a situation where we observe a mode effect, and the mode used is unrelated to Y, either directly (i.e., the variable determines selection into mode) or indirectly (i.e., a variable that causes Y also determines selection into mode). However, there is selection according to a variable X, a variable that is of interest in the analysis, but which in practice has no causal effect upon Y. In this setting, Y* will be spuriously associated with X via mode, M ($X \rightarrow M \rightarrow Y^*$). Stratifying by mode or controlling for mode in a regression will block this path yielding the correct (null) association between X and Y (see data simulations in Figure 4.7). Alternatively, mode effects can be estimated in a regression of Y* upon M to obtain estimates of the counterfactual Y to be used in further analysis (with the same caveats as outlined in Section 4.3).

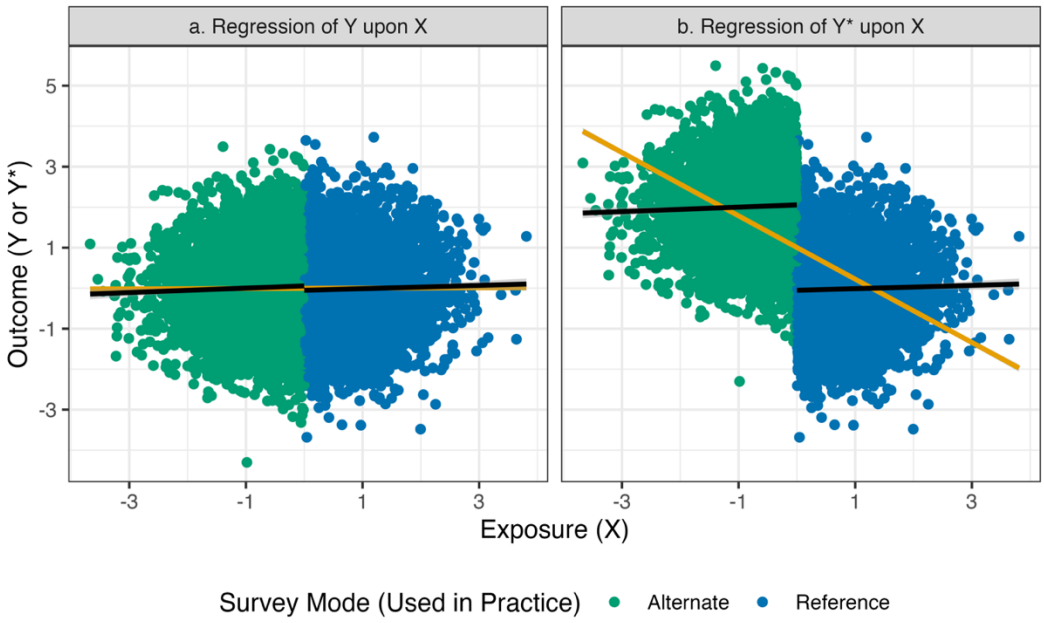

Figure 4.7: Associations between variables X, Y and Y* as calculated via linear regression and reflected in the relationships in Figure 4.6a. Y is subject to a (fixed) mode effect and X determine selection into mode. Orange lines represent regression lines estimated from the total sample. Black lines indicate regression lines when stratifying by survey mode. The colour of the points reflects the mode the survey participant used in practice. Y* is observed value of Y, while Y is the (possibly counterfactual) value that would have been observed had the survey participant used the reference mode. Comparing the panels, there is bias in the association between Y* and X in the total sample which reflects differences in the value of X in each survey mode. An unbiased estimate of the association between X and Y can obtained be stratifying by survey mode (black lines, Panel B). (Note, the lines are based on samples, so gradients are not precisely zero.)

Controlling for mode is also sufficient to block the spurious path from X to Y* via mode where X has a causal effect upon Y (Figure 4.6b). However, X will also need to be controlled for in a regression of Y* upon M to obtain an unbiased estimate of the mode effect, otherwise the path M $\leftarrow$ X $\rightarrow$ Y $\rightarrow$ Y* (or M $\leftarrow$ X $\leftarrow$ Y $\rightarrow$ Y* if Y causes X) will remain open.

4.5 Mode Effects and Collider Bias

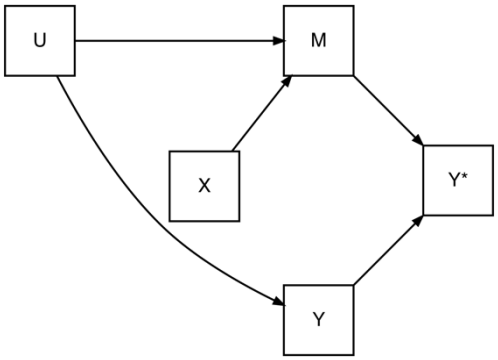

Figure 4.8: Directed Acyclic Graph (DAG) of a situation involving (i) mode effects on a variable Y and (ii) mode selection according to variables X (a variable of interest) and U.

Figure 4.8 extends Figure 4.6a - there is now a further variable, U, that influences selection into mode and is also a cause of Y. The association between X and Y* is biased by mode, but mode is a collider: controlling for mode blocks the path $X \rightarrow M \rightarrow Y^*$, but also opens a path from X to Y* via U ($X \rightarrow M \leftarrow U \rightarrow Y \rightarrow Y^*$). In other words, controlling for M alone does not yield an unbiased estimate of the association between X and Y. This remains true if X has a causal effect upon Y (and vice versa).

Thus, simply adding M to a regression model of Y* on X will not solve the issue of mode effects; controlling for U is also required. Similarly, controlling for U is required to obtain an unbiased estimate of the mode effect from a regression of Y* on M. In practice, though, U may not have been measured or measured poorly. It may even be unknown. Adding M on its own may increase bias in the association between X and Y* relative to the situation where it is not controlled for. The specific degree of bias depends on the relative size of the mode effects and selection effects and the size of the causal effect of U upon Y. To understand the consequences of mode effects for a particular analysis, one therefore needs domain specific knowledge on the variables of interest (including their set of causes) and knowledge of mode selection in the data being used.

It is worth reiterating and expanding on this point. In observational research, to obtain a causal estimate, it is often necessary to adjust for multiple confounding factors. These factors may be unmeasured or measured with error, meaning estimates may

still be biased – often the case as evidenced by the disparity with results obtained using RCTs (see, for instance, Bann et al., 2023), though it should be said that CLS cohorts do contain unusually rich data. Further, due to limitations in current knowledge, the identity of these confounding factors may not even be known to the researcher. Where the researcher intends to use statistical control to account for mode effects, mode selection can add a further set of factors that need to be controlled for.[7] (Sources of mode selection are discussed further in Section 4.8.) Determining the correct set of factors requires knowledge of the data collection procedures and good theories about survey participation. Each of these may be outside the researcher's domain of expertise. In Section 6 we discuss an alternative approach – sensitivity analysis – that only requires appropriate modelling of the mode effect. We believe this would be more robust, as well as straightforward, in situations such as these, which are likely the norm, rather than the exception.[8]

4.6 Mode Effects and Mode Selection

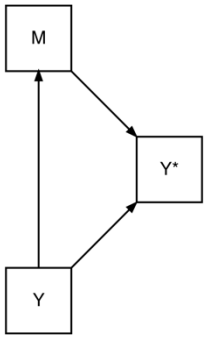

Figure 4.9: Directed Acyclic Graph (DAG) of a situation involving (i) mode effects on a variable Y and (ii) mode selection according to variable Y.

Figure 4.9 represents the situation where Y is both subject to mode effects and a determinant of selection into mode. In this case, M contains information on the latent

[7] In practice, it is possible that these sets of factors overlap entirely requiring nothing extra from the analyst except adding mode to the regression model (assuming these variables have been measured and controlled for already).

[8] Note, we have described controlling for mode via the 'backdoor' method. An alternate approach is to control for the mode effect using the 'front-door' criterion (Cernat & Sakshaug, 2021; Vannieuwenhuyze et al., 2014). In this approach, variables which mediate the effect of mode upon measurement are included as control variables in models. These may include measures of response burden, survey experiences, and susceptibility to social desirability (Vannieuwenhuyze et al., 2014). An advantage is that, as the front-door variables are descendants of mode, controlling for these instead of mode can reduce the extent of collider bias that is induced. However, the method relies upon having information on the full set of mediating variables (otherwise the mode effect will not be accounted for fully), for instance as reports from participants or as survey paradata, but this information is typically not available in surveys. This includes CLS' cohort studies.

variable Y and regressing Y* upon M will not recover the mode effect because of the backdoor path M $\rightarrow$ Y $\leftarrow$ Y*. This would occur, for example, if those with poor mental health were more likely to answer in a survey mode in which individuals tend to understate mental health issues – the association between mode and reported mental health would be appear smaller than is actually the case, as it would be confounded by latent mental health.

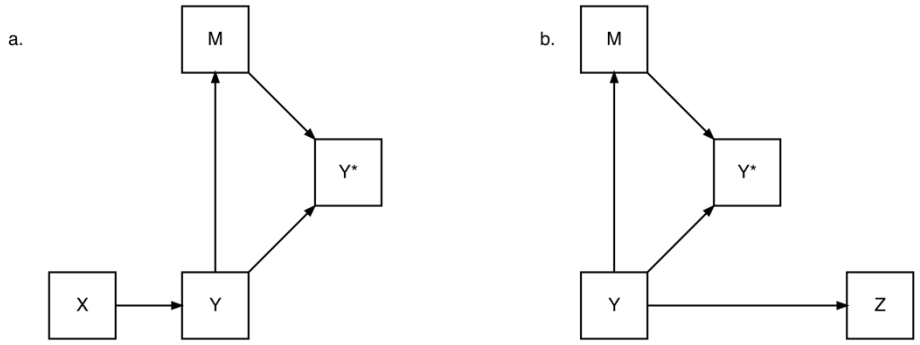

Figure 4.10: Directed Acyclic Graph (DAG) of a situation involving (i) mode effects on a variable Y and (ii) mode selection according to variable Y. In panel (a), Y is caused by a variable X. In panel (b) Y is a cause of variable Z

Figure 4.10a extends Figure 4.9 by including a variable X which is a cause of Y. Controlling for M in a regression of Y* upon X (or alternatively, stratifying by M) would not give an unbiased estimate of the association between X and Y* because M is a descendent of Y; controlling for M implicitly conditions on Y and leads to selection bias as some values of Y are more likely to appear in one mode (if they could be measured) than another.

The consequences of this process can be observed in Figure 4.11, which reflects the relationships in Figure 4.10a with individuals with a positive value of Y using the alternate survey mode and individuals with a negative value of Y using the reference mode.[9] The orange line in the left plot – a regression of (counterfactual) Y upon X – shows the correct causal effect. Regressing (observed) Y* upon X, not controlling for mode, yields an upwards biased estimate (orange line, right plot). This bias reflects the path X $\rightarrow$ Y $\rightarrow$ M $\rightarrow$ Y* in Figure 4.10a – i.e., that values of Y* are partly higher due to the mode effect. Stratifying by M, however, yields bias in the opposite direction,

[9] This is an extreme example used for illustrative purposes only. Section 5 gives examples of factors strongly related to selection into mode in the NCDS and Next Steps.

regardless of whether Y* or Y is used (black lines). Due to the process of selection, high values of Y are not observed in the reference mode and the chance of not being observed increases as X increases. Conversely, low values of Y are not observed in the alternate mode and the chance of not being observed increases as X decreases. (This type of bias is known as bias due to range restriction.) Whether stratifying or controlling for M increases or decreases bias relative to not controlling for M depends on the size of the mode effect and the extent of mode selection (prevalence and the difference in mode according to Y). This is an empirical question.

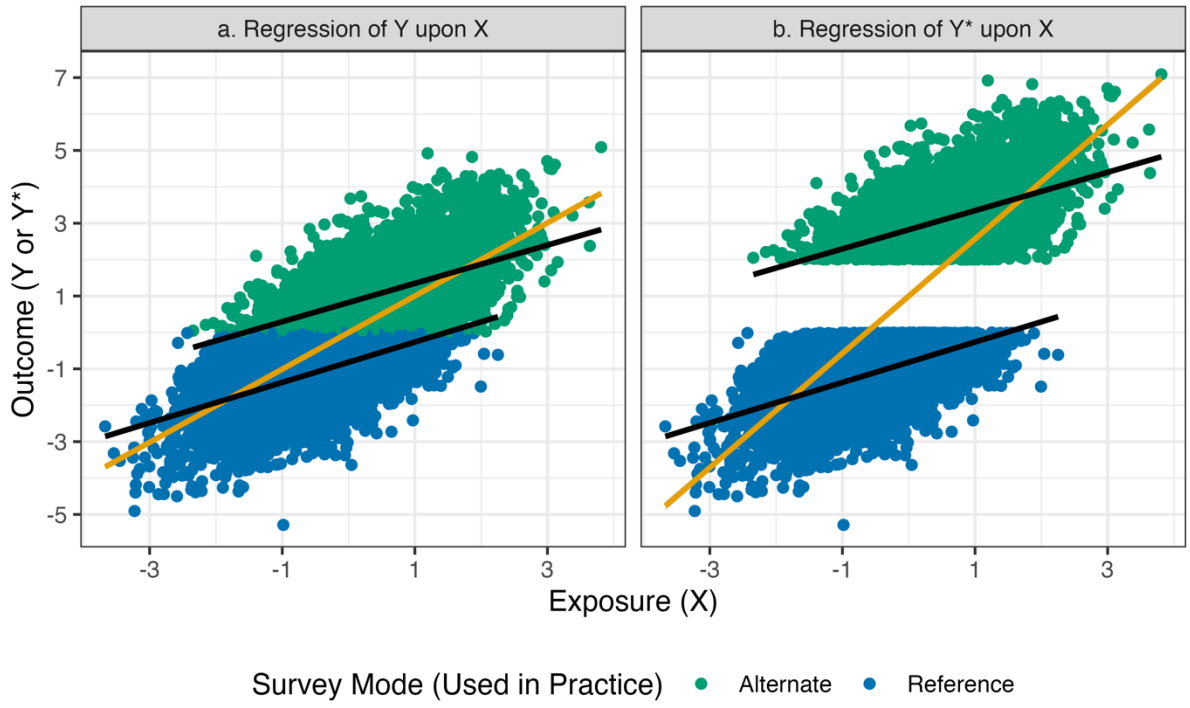

Figure 4.11: Associations between variables X, Y and Y* as calculated via linear regression and reflected in the relationships in Figure 4.10a. Y is subject to a (fixed) mode effect and is also determinant of selection into mode. Orange lines represent regression lines estimated from the total sample. Black lines indicate regression lines when stratifying by survey mode. The colour of the points reflects the mode the survey participant used in actuality. Y* is observed value of Y, while Y is the (possibly counterfactual) value that would have been observed had the survey participant used the reference mode. Comparing the panels, there is bias in the association between Y* and X in the total sample (i.e., not adjusting for M) which reflects increasing chance of a person using the alternate survey mode as Y increases. An unbiased estimate of the association between X and Y cannot be obtained be stratifying by survey mode (used in actuality), however, even if we could observe counterfactual Y (black lines). This is due to range restrictions: we exclude low or high values of Y and Y* from the sample when stratifying by mode.

The range restriction problem does not occur if the aim is to investigate the effect of Y upon another variable, Z, which is a consequence of Y and is neither subject to mode effects nor a cause of mode selection (Figure 4.10b). Here, the restriction in range in Y which is induced by controlling or stratifying by M (which blocks the path $Y^* \leftarrow M \rightarrow Y \rightarrow Z$) does not generate bias because as Y* increases or decreases within a survey mode, no values of Z are more or less likely to be observed (Figure 4.12). Note, differences in the association between Y* and Z between modes could be observed if there are heterogeneous causal effects and these differ on average between the participants in each survey mode. (This is true in other situations where stratifying is sufficient to remove the bias from mode effect, e.g., Figure 4.3b.)

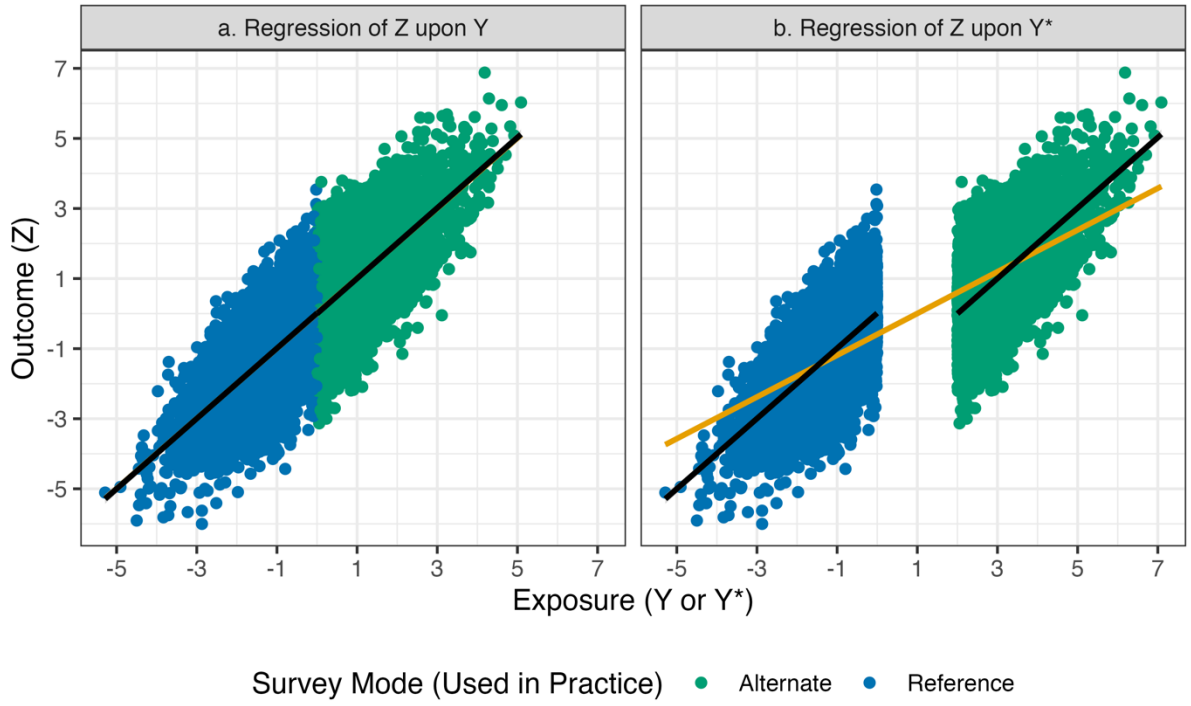

Figure 4.12: Associations between variables Z, Y and Y* as calculated via linear regression and reflected in the relationships in Figure 4.10b. Y is subject to a (fixed) mode effect and is also determinant of selection into mode. Orange lines represent regression lines estimated from the total sample. Black lines indicate regression lines when stratifying by survey mode. The colour of the points reflects the mode the survey participant used in actuality. Y* is observed value of Y, while Y is the (possibly counterfactual) value that would have been observed had the survey participant used the reference mode. Comparing the panels, there is bias in the association between Y* and X in the total sample (i.e., not adjusting for M) which reflects increasing chance of a person using the alternate survey mode as Y increases. An unbiased estimate of the association between X and Y cannot be obtained be stratifying by survey mode (used in actuality), however, even if we could observe counterfactual Y (black lines). This

is due to range restrictions: we exclude low or high values of Y and Y* from the sample when stratifying by mode.

As will be discussed further in Section 6, one option for accounting for mode effects is multiple imputation (Kolenikov & Kennedy, 2014). In this approach, Y is imputed for those in the alternate reference mode using predictive models based on data from those in the reference survey mode. However, where Y is a determinant of mode selection, Y is missing not at random (MNAR) and an assumption required for valid imputation is not met (Rubin, 1987; van Buuren, 2018) – the imputed values of Y will be biased. Where a strong proxy for Y can be included in the imputation model (e.g., a previous measurement of the variable), this bias can be reduced.

The DAGs in Figure 4.9 and Figure 4.10 can also represent situations in which the size of the mode effect depends upon the value of Y. This would occur if susceptibility to the mode effect differs according to Y (for instance, if agreeableness led to greater sensitivity to interview effects for items on personality) or if the mode effect itself makes certain responses more likely. The latter is especially likely for survey items with non-continuous response options – for instance, a binary question or a Likert scale. Here, depending on the value of the underlying latent Y, misclassification can only be in one direction or is otherwise bounded by the limits of the scale – a person responding 'yes' to a binary question in the reference survey mode can only give a false negative in the alternate mode; a person reporting five on a seven-point Likert scale cannot raise their answer by more than two. Midpoint effects and 'straightlining' behaviour (Kim et al., 2019) similarly imply mode effects whose size depend on the answers that would have otherwise been given.

As with mode selection according to Y, heterogenous mode effects according to Y leads to bias in regression coefficients, descriptive statistics, and so on. A further issue is that bias can depend on the type of association parameter – specifically, whether it is 'collapsible' (e.g., absolute risk differences) or 'non-collapsible' (e.g., odds ratios).

4.7 Mode Selection Only

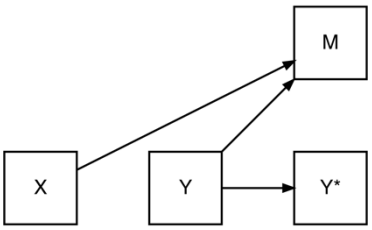

Figure 4.13: Direct Acyclic Graph (DAG) of a situation involving (i) no mode effect on a variable Y and (ii) mode selection according to variables X and Y, where X is not a cause of Y but is of interest in the analysis

The previous examples have shown that appropriately accounting for mode effects with statistical control depends on domain specific knowledge about the substantive (i.e., causal) relationships between analysis variables and the mode selection and mode effects processes. These examples have also shown that systematic differences between modes can reflect mode effects **and** mode selection, and that controlling for mode can induce correlations as mode may be a collider. Figure 4.13 shows a situation where there is mode selection only (i.e., no mode effects). Naively controlling for mode generates an entirely spurious correlation between X and Y* ($X \rightarrow M \leftarrow Y \rightarrow Y^*$). For example, if higher levels of X **or** Y are related to selection into the alternate survey mode, X will be negatively related to Y* among individuals in that mode, as it is less likely to contain individuals with low values of X **and** Y. It is therefore important to be sure that mode effects are present in the data.

4.8 Sources of Selection into Mode

Given the importance of understanding mode selection when attempting to account for mode effects, it is worth briefly discussing the general processes by which selection can occur in mixed mode settings. These are: unit non-response, item non-response, and non-random allocation to offered mode.

Section 3.3 showed how mode can have effects on unit and item non-response. Potential participants may be uncontactable in some modes or could refuse some types of survey, and participants may be unwilling or unable to answer certain items in particular modes – for instance, due to lack of motivation, discomfort disclosing information to an interviewer, or insufficient information or inability to ask clarifying questions (i.e., to understand complex items). Importantly, non-response can be due

to the individual (e.g., unwillingness to discuss sensitive topics with an interviewer) or arise from the survey design itself (e.g., lack of usable contact details), but in either case can lead to differences in the characteristics of individuals providing data in each mode. Item and unit non-response generate selection effects even where participants are randomly allocated to mode *offered*.

Surveys do not typically allocate individuals to mode randomly – for instance, in sequential mixed mode designs, some modes are offered only after non-response. This non-random allocation is another source of selection effects, and again can reflect individual decisions or survey design (or their combination). In some cases, selection is explicit. For instance, participants may be automatically assigned to a mode because of information held about them (e.g., NCDS Sweep 9 participants were allocated to the web mode if no telephone number was held for them) or may choose a mode because they prefer it (as with the MCS Sweep 6 Time User diaries). In other cases, selection is implicit. For instance, in sequential mixed mode designs, individuals who are difficult to contact or who take longer to agree will only appear in later offered modes.

In each case, non-random allocation is likely to lead to differences between participants in each mode. These differences will also extend beyond the characteristics directly determining allocation. For instance, participants allocated to web survey because they do not have a telephone will also differ from other participants according to the characteristics that led them to not have a telephone, that are consequences of not having a telephone or that share a common cause with not having a telephone. The number of factors directly contributing to allocation can also be many. In sequential mixed mode designs, several factors are likely to proximally influence unwillingness or inability to answer a survey at the first contact attempt (e.g., curiosity, family and work demands, etc.). Each of these will also be related to a range of other characteristics. For instance, in their analysis of Health Survey for England data, Boniface et al. (2017) show that participants who required more contacts before participating had higher reported alcohol consumption. Where selection effects are strong enough, finding comparable sets of individuals in each mode (necessary to estimate mode effects unbiasedly) may be infeasible.

4.9 Joint Mode Effects

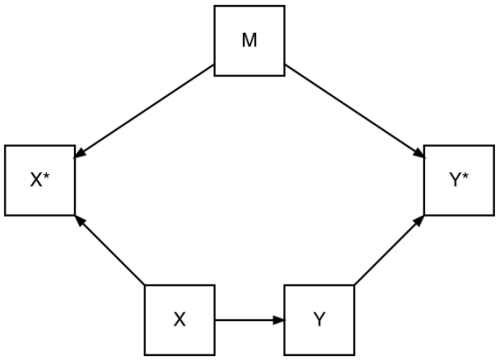

Figure 4.14: Directed Acyclic Graph (DAG) of a situation involving mode effects upon two variables, X and Y, both of which are of interested in substantive analysis.

Bias can also arise where there are mode effects only. Figure 4.14 displays the situation where two variables of interest (X and Y) in the substantive analysis are each subject to mode effects but are not causes of mode selection. The association between X* and Y* is biased by the backdoor path $X^* \leftarrow M \rightarrow Y^*$. Conditioning upon M is sufficient to block the association arising via mode effects. It is also sufficient where there is additionally mode selection according to a variable that is not X or Y, even if it is a cause of either variable.

4.10 Mode Effects in Longitudinal Analysis

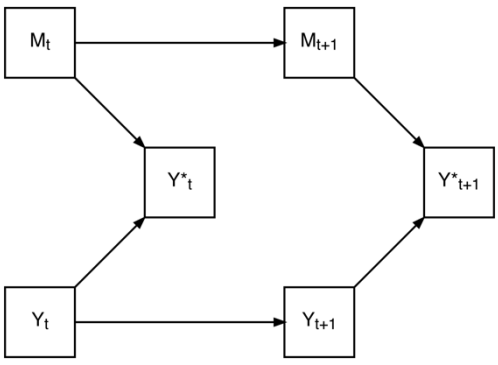

Figure 4.15: Directed Acyclic Graph (DAG) of a situation involving mode effects upon a variable Y which is measured over two time periods, t and t+1.

A special case of the previous example is using longitudinal repeated measure data. Mode effects can induce additional correlation between measurements if modes are correlated temporally ($Y_t^* \leftarrow M_t \rightarrow M_{t+1} \rightarrow Y_{t+1}^*$; Figure 4.15) – for instance, if participants who answer via face-to-face interview at one time point are more likely to

answer via face-to-face interview at another . However, even if mixed mode was used in only one sweep, or different modes were used across sweeps, estimates of change over time (which by construction reflect the strength of the relationship between adjacent measurements of Y) will still be biased as temporal change will be confounded with mode switching. Bias can also arise from more substantive concerns. Specifically, measurements in one mode may be more sensitive to changes that occur over time. For instance, psychological distress increases from early- to mid-adulthood (Bell, 2014; Gondek et al., 2021), and modes that are less likely to understate mental health problems (e.g., anonymous surveys) may better capture this change.

4.11 Mode Effects in Mediation Analysis

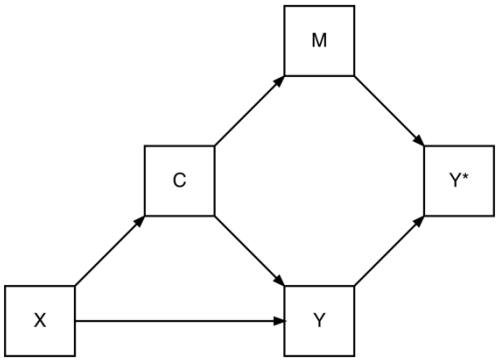

Figure 4.16: Directed Acyclic Graph (DAG) of a situation relevant to mediation analysis where there is (i) a mode effect on a variable Y and (ii) mode selection according to a mediator (C) of the relationship between variables X and Y.

Another form of analysis that is widely used with longitudinal data is mediation analysis – i.e., where an attempt is made to decompose the total effect of a variable upon another into direct and indirect (mediated) effects. Figure 4.16 shows one such set-up. Here the effect of X upon Y is partly mediated by a variable, C, which is also a source of mode selection. Without controlling for mode, the indirect effect of X upon Y* via C is biased by mode selection, $X \rightarrow C \rightarrow M \rightarrow Y^*$. The direct effect (obtained by conditioning upon C) is unbiased. However, it is typical in mediation analysis to express direct and indirect effects as proportions of the total effect – given the indirect effect is biased, the overall proportion represented by the direct effect will be, too. Nevertheless, in Figure 4.16, mode effects can be accounted for by conditioning upon mode and mode effects can be estimated by regressing Y* upon M and C.

Figure 4.16, of course, displays just one mediation scenario – direct effects would be biased if X was a direct cause of mode selection (i.e., not mediated by C) and issues would also arise if the exposure or mediator exhibited mode effects. Further, as shown in Section 4.5, where there are other causes of mode selection that also influence the variables of substantive interest, conditioning upon mode could lead to collider bias. Mediation analyses are particularly difficult with observational data (but even with RCTs) as, for the assumptions to be met, it is necessary to control for confounders of the exposure-mediator, exposure-outcome, and mediator-outcome relationships, a large ask. Mode selection adds another set of confounders to this.

4.12 Mode Effects in Moderation Analysis

A final type of analysis we will explore is moderation analysis – i.e., examining whether an effect of a variable differs across strata of another variable, either using stratification or by adding an interaction term to a regression model. The principles outlined in Sections 4.3 - 4.7 apply to this form of analysis, too. For instance, where there is a mode effect only, a regression of Y* upon X interacted with a moderator variable D will not yield biased regression coefficients, but a regression of Z upon Y* interacted with D will; both main effects and interaction terms will be biased, in this instance. Controlling for mode will remove the bias, but an interaction term between mode and the moderator variable is additionally required.

Mode selection potentially adds a further source of bias, but this depends on the variable determining mode selection and the quantity of interest (main effects or interaction terms only). For instance, where the moderator variable D is related to mode selection, a higher proportion of observations in the alternate survey mode will be observed in one level of the moderator compared with other levels. Where Y* is being used as an outcome measure, mode selection on D but not X will lead to biased coefficients for main effects but not interaction terms (mode effects on both will bias both coefficients). Controlling for mode, interacted with D and X will restore the unbiased estimates. These examples again are not intended to be exhaustive; other consideration may arise, such as mode effects in multiple variables and collider bias arising from conditioning upon mode.

4.13 How Important are Mode Effects and Mode Selection in Practice?

The previous sections show a range of scenarios in which mode effects can lead to bias in analyses of survey data. Among these are scenarios in which there may be no set of variables available to remove this bias using statistical control – for instance, where there are unobserved factors that are a source of collider bias when conditioning on mode, or where a variable subject to mode effects is also a determinant of mode selection. The extent of this bias is an empirical question and depends on the proportion of individuals in the alternate mode, the size of the mode effect, and the strength of mode selection (if applicable). In practice, none of these may be sufficient to cause a level of bias that is material to results, especially where the aim of the analysis is to obtain an overall direction of association, rather than accurately determine a parameter value to be used in decision making – the former being typical of most social science research, while the latter more important in areas such as economic evaluation and decision modelling. In the next section, we provide empirical evidence on mode effects and mode selection in the NCDS and Next Steps.

5 An Empirical Assessment of Mode Effects and Mode Selection

As discussed in Section 4, the extent to which mode effects will lead to bias in results depends on the size of the mode effect, the proportion of participants in each mode, the strength of mode selection, and the association variables exhibiting mode effects have with mode selection factors. In this section, we provide empirical evidence on mode effects and mode selection using data from the NCDS and Next Steps. We also give an overview of mode effects estimates obtained from mode experiments embedded within four major health and social science surveys. We provide this evidence so that analysts can make informed judgements as to the potential bias in their own analyses of CLS data and can posit sensible parameters for mode effects when simulating data in sensitivity analyses – an approach that we recommend for handling mode effects (see Section 6). Specifically, we provide overviews of: (1) mode effect estimates from the survey mode experiment in Sweep 9 (age 55y) of the NCDS (Goodman et al., 2022); (2) mode selection estimates from Sweep 9 of the NCDS; (3) mode selection estimates for Sweeps 4-8 of Next Steps; and (4) an overview of mode effects research using data from the United Kingdom Household Longitudinal Study (UKHLS), the Health and Retirement Study (HRS), the European Social Survey (ESS), and the third round of the National Survey of Sexual Attitudes and Lifestyles (Natsal-3).

5.1 Mode Effects in Sweep 9 (Age 55y) of the NCDS

Sweep 9 (age 55y) of the NCDS embedded a survey mode experiment in which eligible cohort members were randomized to sequential mixed mode (web then telephone interview) or telephone-only data collection. Goodman et al. (2022) analyse the results of this experiment to estimate mode effects. Here, we provide a summary of their results, as well as referring readers to a table of their mode effect estimates (Supplementary Table 1), which we have extracted from their paper and to which we have added information on source variables and standard deviations, the latter so that Cohen's *d* effect sizes can be calculated (useful for simulating mode effects for similar survey items).

It is worth first noting that Sweep 9 of the NCDS followed design principles to reduce mode effects, including predominantly collecting factual information (Brown, 2016). Nevertheless, some sensitive and subjective questions were still asked (for instance, on health and mental wellbeing) and there were significant differences between modes. Importantly, telephone respondents answered items aloud to an interviewer – typically with yes or no answers to protect from being overheard – while the web survey was entirely anonymous. This context should be considered when applying results to other studies. For instance, while Sweeps 5-8 of Next Steps were also designed to reduce mode effects (Calderwood et al., 2017; Department for Education, 2011), the survey contained a more extensive set of sensitive questions, including on topics such as drug use, sexuality and sexual histories.

As participants in the NCDS were randomized to mixed mode rather than web alone (~ 25% of participants in the mixed mode arm completed the survey via telephone), Goodman et al. (2022) calculate the 'complier-average causal effect' (CACE) using experimental arm as an instrument for whether the participant responded via web. This is the mode effect among those who ultimately answered via web survey and assumes that telephone responses in the mixed mode arm were not influenced by the offer of web.[10] As the offer of web also had an effect on unit and item non-response, they imputed missing values, assuming missingness at random (Rubin, 1987).

Of the survey items examined in their analysis, Goodman et al. (2022) find one-third exhibit statistically significant mode effects at the 5% level; the experiment was powered to detect, with 80% power, a 0.08 SD difference for continuous variables and maximum 3.9 percentage point difference for binary variables. Most of the statistically significant mode effects were small. Mode effects for the two continuous items that exhibited the largest mode effects (items 3 and 6 of the CASP-6 measure on quality of life) were approximately 0.23 SD – participants in the (anonymous) web survey mode reported lower wellbeing on average. This is comparable to the association between current unemployment and mental distress found in longitudinal studies (Paul & Moser, 2009).

[10] In sequential mixed mode settings, this assumption may not apply as the sequence of offered modes may determine when a participant responds and this could influence responses (e.g., there is some [inconsistent] evidence that depressive symptoms mental health shows seasonal fluctuation [Øverland et al., 2020], so the date of survey completion can influence responses).

Items that exhibited statistically significant mode effects were typically those that reflected sensitive topics, particularly where a subjective assessment was required. This included items on physical and mental health, alcohol use (units consumed though not frequency), physical activity, financial stress, likelihood of working at age 66, and eating in pubs or restaurants weekly, the latter two possibly signifiers of wealth. In each case, participants in the telephone survey mode gave more favourable responses (e.g., reported fewer health problems and lower financial stress). Primacy and recency effects did not appear to have a large influence on responses – the CASP-6 measure contains positively and negatively worded items, but in each case, responses in the telephone survey tended towards higher levels of wellbeing.

Items that did not exhibit statistically significant mode effects typically related to easy-to-recall factual information, such as whether the participant was employed and in a professional occupation, hours of work, number of qualifications and relationships since last survey, smoking and frequency of alcohol use. However, differences were observed for whether a person reported having a longstanding illness or being classified as disabled, for specific health conditions such as backache and problems with hearing (though these may be undiagnosed), and voting Liberal Democrat in the 2010 General Election.

Mode effects were also observed on item non-response (see Table 7 of Goodman et al., 2022). Participants in the telephone mode were less likely to provide valid responses to items on expected value of property, remaining mortgage amount, and gross weekly income, but were more likely to report employer provided pension type. Some of these effects were sizable. For instance, the telephone survey was estimated to reduce response rates for value of property by 6.3 p.p..

5.2 Mode Selection in Sweep 9 (Age 55y) of the NCDS

Goodman et al. (2022) also examined selection into web or telephone survey among those in the mixed mode experimental arm. Again, we provide a summary of their results, as well as referring readers to a table of their estimates (Supplementary Table 2) that we have extracted from their paper and to which we have added information on source variables and standard deviations for continuous variables so that Cohen's *d* effect sizes can be calculated. Note, as not all individuals were assigned to the mixed

mode arm of the survey mode experiment, selection into mode among all participants will differ to that in the mixed mode experimental arm.[11]

Goodman et al. (2022) report a number of differences in background characteristics according to survey mode used. Background characteristics were measured at age 50 or earlier, when only single survey modes were used, so do not exhibit mode effects. Predictably, there were large differences in computer use according to survey mode: 43% of telephone respondents used a computer at home 2 or more times a week at age 50 compared with 74% of web respondents. Participants also differed according to socioeconomic and educational characteristics: more advantaged individuals were more likely to respond via web. 52% of web respondents were employed in a professional or managerial occupation at age 50 compared with 32% of telephone respondents. There were also important differences according to physical and mental health, health behaviours, and cognitive ability. Telephone respondents were 13 pp. more likely to have been smoking at age 50 and measures of cognitive ability were > 0.3 SD higher on average among web respondents than telephone respondents. It is likely that these differences would have been larger if contemporaneous characteristics and behaviours were used. However, it was of course not possible to examine this robustly given responses at age 55y may have exhibited mode effects.

5.3 Mode Selection in Sweeps 5-8 of Next Steps

Supplementary Table 3 shows the characteristics of participants in each survey mode in Sweeps 5-8 (ages 17/18y – 25y) of Next Steps. These sweeps adopted a sequential design with web offered before telephone and then face-to-face interview. Characteristics in this table were measured at Sweep 4 (age 16/17y) in which only face-to-face interview was used, and thus should not be biased by mode effects.[12]

There were differences in mode used according to most of the variables examined. Compared with web respondents, face-to-face respondents were more likely to be male, to be from socioeconomically disadvantaged backgrounds, to have younger

[11] 12.5% of total participants were eligible but randomly assigned to the telephone-only arm. 86.8% of 441 participants ineligible for the experiment participated in the web survey, with the remainder responding by telephone. The main reasons for ineligibility were (a) being an emigrant (responded by web), (b) not having a telephone number on file, and (c) a dress rehearsal case. Reasons for ineligibility are stored in the variable `N9MODEXL`.

[12] Exceptions to this are sex and maternal and paternal age at birth, each time-invariant and an easy-to-recall characteristics (parental age at birth was derived using birth dates for parents and cohort members).

parents, special educational needs, to have been in care, been NEET at age 16/17, have poorer health and to display risk factors such as carrying a knife, being a victim of violence or trying cannabis. Some of these differences were large. For instance, 29.1% of face-to-face respondents at Sweep 5 had parents in the highest socioeconomic occupational class, compared with 53.5% among web respondents. An exception was mental health at age 16/17 (GHQ-12 scores), which showed little consistent difference according to survey mode. However, this does not preclude contemporaneous mental health being a determinant of mode selection. Telephone respondents also differed from web respondents according to most of the examined factors and in the same direction as face-to-face respondents, but differences were typically smaller in size.

5.4 Mode Effects in Other Studies

Mode effects have been estimated in several other surveys that have embedded mixed mode experiments. Here we discuss results from four major health and social science surveys covering similar topics to CLS' cohort studies: the United Kingdom Household Longitudinal Study (UKHLS, also known as Understanding Society), the US Health and Retirement Study (HRS), the European Social Survey (ESS), and the third round of The National Survey of Sexual Attitudes and Lifestyles (Natsal-3).

The UKHLS has embedded multiple mixed mode experiments. Wave 8 of the UKHLS survey included a mode experiment in which households were randomized to face-to-face then web or web then face-to-face sequential mixed mode designs. Clarke & Bao (2022) use data from this experiment to examine mode effects (web compared with face-to-face) on the means and variances of survey items. They find statistically significant (at the 5% level) differences in 13% of 166 variables examined. Among variables exhibiting statistically significant mean level mode effects were: pay period, commuting distance, and work location. Among variables exhibiting statistically significant variance level mode effects were: cigarettes smoked, commuting distance and gross income from second job. Variables not exhibiting statistically significant mode effects included physical and mental health as measured by the SF-12. Unfortunately, the authors only report a selection of results and is not clear how these were chosen.

In addition to the main survey sample, the UKHLS maintains an ‘Innovation Panel’, which is used to test novel survey methodologies. Wave 2 of the Innovation Panel embedded a mode experiment in which participants were randomized to sequential mixed mode (telephone then face-to-face interview) or single mode (face-to-face only) designs. Cernat (2015) examined responses to the 12-item Short Form (SF-12) questionnaire, a widely used measure of mental and physical health, finding that only one item (‘felt calm and peaceful’) differed by mode. Telephone respondents reported lower frequency of feeling calm and peaceful, a difference that is consistent with social desirability or optimism bias influencing responses (Cernat, 2015). The author also found that mode effects biased longitudinal estimates of change (Waves 1, 3 and 4 used face-to-face interviewing only). Within-individual variance was greater in the mixed mode group for four of the SF-12 items. Each of these four items tapped mental health.

Waves 5 of the Innovation Panel embedded a similar experiment to Wave 2. Participants were randomized to web then face-to-face interview sequential mixed mode or face-to-face only designs. In unpublished work, Jäckle et al. (2016; cited in Jäckle et al., 2017) find only 4-9% of the 479 items they test exhibited mode effects. Items that were more likely to exhibit mode effects required responses on rating scales or had five or more response categories (Jäckle et al., 2017).

Like the UKHLS, the HRS has also included multiple mode experiments. The 2018 sweep of the HRS embedded a mixed mode experiment with participants randomized to telephone only or web then telephone sequential mixed mode designs. Ofstedal et al. (2022) perform an intention-to-treat analysis of differences in responses to items on health, expectations about the future, and financial assets (79% of mixed mode participants responded via web). The authors find evidence of small mode effects on item means, with telephone respondents tending to give more optimistic and socially desirable answers. Effect sizes for measures of health were less than 0.1 SD, including for measures of depressive symptoms and life satisfaction. An ‘optimism score’ calculated from responses on expectations about future health, finances and the economy showed a stronger mode effect, but the effect size was still only ~0.2 SD. Differences in responses regarding financial assets were especially small.

Domingue et al. (2023) use cognitive test data from the same HRS experiment and find scores have higher mean (0.3 SD) and lower variance (~ 0.4 SD) among web respondents. (See Al Baghal, 2019, for similar results in the UKHLS.) The authors suggest this may be due to web respondents having access to their computer to work out correct answers. The authors also show that mixing modes between sweeps biases longitudinal estimates of change, particularly when comparing web responses with face-to-face responses from prior sweeps – in this case, mode effects are sufficient to make cognitive ability appear to improve over time, contrary to other research on cognitive ageing.

Cernat et al. (2016) use data from the 2010-2012 sweeps of the HRS, exploiting a feature of the survey in which participants were randomized to telephone then face-to face interview or face-to-face then telephone interview in successive biennial sweeps, with a web survey collected from every participant in the intervening year. The authors compare reported depressive symptoms (CES-D depression scale) and physical activity by survey mode finding that participants reported higher depressive symptoms but also more physical activity in the web mode. Responses did not differ between telephone or face-to-face modes, suggesting social desirability was an important influence on mode effects (the CES-D contains positive and negative coded items, so differences cannot be explained by primacy or recency effects). Analyses using latent variable modelling suggest a sizeable mode effect (~ 0.4 SD) for depressive symptoms. The size of the mode effect also varied according to the level of the latent depression score.

Evidence that social desirability concerns generate mode effects has also been found in experiments embedded within the ESS. Jäckle et al. (2006) find less socially desirable reporting on political attitudes and beliefs in telephone versus face-to-face modes. They also find evidence of primacy and recency effects, though for fewer items than are influenced by social desirability. (Results of other experiments within ESS are summarized in Villar & Fitzgerald, 2017.)

Mode effects in Natsal have been examined via retesting in a different mode. Burkill et al. (2016) use data from a web follow-up of Natsal-3 participants, comparing responses to those from the main face-to-face interview which used a self-completion component for particularly sensitive questions. The authors find differences for many

survey items between the web and main interviews: participants gave more socially desirable responses in the main interview, even though self-completion was used for many of the questions. This suggests that physical presence alone is sufficient to induce interviewer effects. Mode effects also differed between males and females for several items. For instance, males reported fewer sexual partners over the last twelve months in the web mode, while for females, the opposite was true (though, mode effects for lifetime partners did not differ by gender).

5.5 Summary

On balance, the literature above suggests social desirability is the primary driver of mode effects, though primacy and recency effects can also influence responses. Interviewer effects may arise even in self-completion modules where the interviewer neither sees nor hears the answers. However, mode effects, even for socially valent survey items, are typically small – 0.2 SD or less is typical for effects upon item means. There is some evidence for distributional effects; for instance, effects on variance and on the likelihood that extreme values are selected (primacy and recency effects). Mode effects are also not always consistent between individuals. What is considered socially desirable can vary across groups, and this can influence mode effects: evidence from Natsal has demonstrated differences in mode effects on reports of sexual activity by gender. Other characteristics that could determine relevant social norms include generation (i.e., birth cohort), country of origin, and social and cultural background.

6 Methods for Handling Mode Effects

Several methods are available for handling mode effects, each with advantages and disadvantages. In this section, we describe four approaches in detail, highlighting their positives and drawbacks (Table 6.1). Recommendations are provided in Section 7. Ultimately, the appropriateness of a particular approach depends on the analysis being carried out, the data that are available, and assumptions about mode effects and mode selection. This can be captured using a DAG (Section 4).

Table 6.1: Approaches for Handling Mode Effects

Approach	Description	Assumptions
Statistical Control	Mode effect is accounted for either by (a) incorporating mode into the main analysis model directly or (b) estimating the mode effect and using this to predict counterfactuals for those observed in the alternate mode.	Analysis variables are not related to mode selection or mode selection is correctly accounted for, either by adding relevant causes of selection into statistical models as controls or exploiting exogenous variation in mode (i.e., as an instrumental variable)
Multiple Imputation	Observations for variables plausibly suffering mode effects are deleted for those in the alternate mode and values imputed using data from the reference mode to obtain counterfactual value	Data are missing at random: conditional on the used data, answering in the alternate mode is not informative about the value of the variables to be imputed. This is equivalent to requiring that mode selection is correctly accounted for.
Sensitivity Analysis	Size of the mode effect is posited and used to 'correct' observed data with main analysis then run with this amended data. The process can	Mode effects have been correctly modelled. Plausible or extreme values for mode effect can be assumed: if results are robust to extreme mode effect, this would

	be repeated across a range of posited mode effect sizes.	suggest mode effects do not drive results.

6.1 Accounting for Mode Effects with Statistical Control

As detailed in Section 4, where the aim is to obtain an estimate of association (e.g., a beta coefficient from a regression model, or, under a set of assumptions, a causal effect), it may be possible to account for mode effects using control variables. In the most straightforward situation where there is no relevant mode selection, this would simply involve adding an indicator variable for mode to the substantive model or, alternatively, stratifying by mode. However, where there is mode selection according to a relevant variable, this approach may not be sufficient and could even increase bias. In this case, a larger set of control variables, or a method which exploits exogeneous variation in mode (e.g., instrumental variables), may be required. It is possible that this set of control variables does not exist in the data or could not exist in the data (for instance, where a variable, Y, is both subject to mode effects and a source of mode selection; Figure 4.9). It may nevertheless be worthwhile running models with available controls if this would be expected to substantially reduce bias relative to not taking such a measure; in practice, observed variables may well approximate selection effects. The reduction in bias would depend on the relative size of the mode effects and the collider bias induced by conditioning upon mode (the latter of which is a function of the extent of mode selection and the degree of confounding, both empirical questions).

Statistical control may be achieved using an instrumental variable approach where there exists a variable that determines selection into mode that is also not related to variables exhibiting mode effects. Experimental arm in the NCDS Sweep 9 mixed mode experiment is one such variable, at least among eligible participants and subject to assumptions about unit and item non-response. Other variables could, in theory, be identified by studying data collection or mode selection in detail, though we are not aware of any in CLS' cohorts.[13]

[13] The COVID-19 pandemic led to the use of video interviewing for some CLS cohort members. However, surveys were not issued randomly over time, so this is unlikely to work as an instrument (further, responses may differ because of the pandemic – for instance, effects on mental wellbeing).

Another situation in which the statistical control approach is less than ideal is where mode effects are heterogeneous. Where possible, the heterogeneity should be modelled, e.g., by including interaction terms in the substantive model. However, it is possible that the available data are not sufficient to do this appropriately (for instance, where the size of the mode effect is related to the underlying value of the variable that is subject to mode effects [see Section 4.6]). Heterogeneous mode effects alter the variance of the variable, which introduces attenuation bias and cannot be accounted for with a single fixed effect term.

Finally, a further issue with statistical control is that it can change the quantity being estimated. Obtaining a causal estimate is just one aim of analysis – one may be interested in calculating an association for purely descriptive reasons (e.g., to measure health inequality). Adjusting for causes of mode selection to avoid collider bias may change the interpretation of the estimate being produced.

6.2 Estimating the Mode Effect and Using the Predicted Counterfactual

An issue with the above approach is that it is most useful for analyses looking at associations. The method can be straightforwardly used to obtain descriptive statistics for the counterfactual mode where there are mode effects but not mode selection – for instance, one could use linear regression, generalized linear modelling, or GAMLSS (Rigby & Stasinopoulos, 2005) to model the distribution of the variable and make inferences from the model estimates (e.g., the intercept would be the counterfactual mean where all covariates are set to zero). Another approach is to estimate the mode effect directly and use this to predict the counterfactual value for those in the non-reference mode. These values can then be used to generate descriptive statistics or in other forms of analysis. Below is pseudocode showing this process. Uncertainty in the estimate of the mode effects should be propagated, for instance using bootstrapping.[14]

```
regress y m ${control_variables}
```

[14] It is worth noting that there may be differences in responses between modes that are neither due to mode selection nor mode effects, but rather in the underlying latent construct a survey item is intended to capture. Specifically, in the sequential mixed mode design, fieldwork for later modes might start only after fieldwork for earlier modes has ended. Differences between modes may be a product of the date of data collection (e.g., economic activity may reflect the seasonality of work).

```
generate y_counterfactual = y - _b[m] * m
mean y_counterfactual
sd y_counterfactual
…
```

It is not necessary for the mode effect to be estimated using standard regression – other approaches, such as matching, inverse probability weighting, measurement equivalence testing (Cernat, 2015; J. J. Hox et al., 2015; Sakshaug et al., 2022), or instrumental variables regression (Goodman et al., 2022) could be used. More complicated techniques that are capable of modelling effects upon other moments or parameters of the variable distribution, such as quantile regression, generalized method of moments (Clarke & Bao, 2022) or GAMLSS, could also be adopted. If one is willing to make the assumption that the mode effect is rank preserving (as assumed in the above pseudocode), the observed value can be mapped onto the corresponding value of the predicted counterfactual distribution. For instance, if the data are normally distributed with mean = 0, SD = 1 in the estimated counterfactual (reference survey mode) distribution and mean = 2, SD = 3 in the observed (non-reference mode) distribution, a value of 5, which is one standard deviation above the mean of observed distribution, would be a value of 1 in the counterfactual distribution.

This approach – and the approach in Section 6.1 – makes use of the information contained within the observed values of the alternate mode. The mode effect is subtracted from the observed values, but information from the observed values is otherwise retained; more formally, this process uses both the predictions *and* residuals implied by the mode effect estimating model (`regress y m ${control_variables}` above). This works straightforwardly for continuous variables but does not do so for categorical variables (Kolenikov & Kennedy, 2014). With categorical variables, the outcome (e.g., a 0 or 1) is different to the estimand (typically, a probability or odds ratio). Subtracting the mode effect from the observed value changes the scale and does not produce a quantity that is interpretable or appropriate for analysis. Predicted probabilities implied by the mode estimating model can be used to calculate descriptive statistics (the weighted sum of predicted probabilities equals [expected] prevalence; Kolenikov & Kennedy, 2014), but these predictions discard information on observed outcomes, increasing measurement error, reducing precision and inducing attenuation bias in relevant situations.

Given this, Kolenikov & Kennedy (2014) introduce a latent variable method that uses information on the observed outcome (Y*) to determine the distribution of residuals from logistic regression models estimating mode effects. These residuals are then used to simulate counterfactual values of Y with repeated draws made to propagate uncertainty in the estimate of the mode effect and the residuals. (This method also works for mode effects estimated for categorical variables with the multinomial logistic regression model.) However, while this method makes use of the information contained within the observed data, to our knowledge, it is not possible to implement it straightforwardly using out-of-the-box commands in the major statistical programming languages.

As with the statistical control approach (Section 6.1), the validity of the counterfactual approach hinges on the available data and whether there is mode selection – it may not be possible to accurately estimate mode effects if sources of mode selection are not accounted for. Heterogeneous mode effects also generate problems as the prediction of the counterfactual will not be accurate if there is severe heterogeneity – though, as noted, it may be possible to account for this using a modelling approach that allow for effects on multiple parameters of the distribution, subject to the assumption that the mode effect is rank preserving, which may be implausible in practice.

Heterogeneity may also be important when estimating mode effects using sequential mixed mode experiments (such as in Sweep 9 of the NCDS) as the mode effect that can be estimated is the complier average causal effect among participants eligible for the experiment (Goodman et al., 2022). This is not necessarily the same as the mode effect that be observed among participants ineligible for the experiment or non-compliers.

6.3 Multiple Imputation

An alternative approach that is straightforward to implement for a wider variety of variable types is multiple imputation (MI; J. Hox et al., 2017; Kolenikov & Kennedy, 2014). In this approach, values of variables hypothesized to exhibit mode effects are artificially set to missing for individuals in the alternate survey mode(s). Values are then imputed using predictive models based on data from those in the reference survey mode; predictions are made by applying the predictive models to data (e.g.,

covariates) that has been retained for those in the alternate survey mode. For instance, a prediction for Y may be made based on the participant's sex and age using statistical associations identified among those in the reference mode. The imputed values represent predicted counterfactuals for those in the alternate mode and can be used to generate descriptive statistics or analysed in substantive regression models. Multiple imputed datasets are generated by this procedure. Estimates need to be pooled, e.g., using Rubin's (1987) rules, to obtain standard errors that account for uncertainty inherent in the imputation models; less accurate predictions lead to greater between-imputation variability and hence increase uncertainty.

MI has several advantages over the approaches described in Sections 6.1 & 6.2. Importantly, there is easy-to-use and thoroughly documented functionality for implementing MI in each of the major statistical programming languages (e.g., the `mice` or `jomo` packages in R and the `mi` commands in Stata). This includes simple commands to obtain appropriate standard errors for substantive models. MI is also widely-adopted method with ample training and guidance available (e.g., see van Buuren, 2018). Researchers may already be using MI to handle item-level and unit-level missingness in their analysis. The imputation of reference survey mode counterfactuals can be bundled into this step.

MI also generates counterfactuals that are on the correct scale – for instance, imputing binary variables rather than predicted probabilities. Most common distributions and variable types can be imputed, including categorical variables, truncated distributions, and multilevel data. One particularly useful imputation algorithm is predictive mean matching (PMM). PMM uses donor observations as imputation values, which constrains imputed values to be drawn from the same set of values as the observed data. This is particularly helpful when imputing survey items with response scales: imputed values cannot sit outside the observed range or take non-integer values where this is not permissible.

However, MI has two important limitations that may make it unsuitable given the substantive analysis being performed. First, MI (in this context) is wasteful. It does not use information from the observed values in the alternate mode. The degree of this wastefulness depends on the proportion of participants in the alternate survey mode and the predictive accuracy of the imputation models. It is greater where multiple

variables are hypothesized to be subject to mode effects as each will be artificially set to missing in the alternate survey mode; observed values in the alternate mode for one variable will thus not be used for the predictions of others. Second, MI assumes the data are 'missing at random' (MAR), which means that missingness should be independent of the (counterfactual) value of Y, conditional on the covariates used to generate the imputed values. This will not be the case where Y is a source of mode selection. (The MNAR assumption will also be violated if Y is caused by another variable which is a cause of mode selection, if this variable is not included in the imputation model.) Imputed values of variables that are 'missing not at random' (MNAR) are biased. For example, if Y is positively related to selection into the alternate mode, imputed values of Y will be smaller than they should be.

Both limitations can be reduced by incorporating highly predictive information into imputation models. CLS's cohorts have the advantage that they each contain extremely rich data collected over the life course and often in single survey modes. This includes hundreds, if not thousands, of variables collected at each sweep and the repeat measurement of many traits, including mental and physical health, socioeconomic outcomes, and BMI, that show strong within-person correlation over time (see, for example, Norris et al., 2020). However, the predictive power of this data should not be overestimated; prediction algorithms developed in other surveys with extensive data collection have performed surprisingly poorly (Salganik et al., 2020; Seligman et al., 2018). Thus, some bias and inefficiency when using MI may remain, though as sample sizes in CLS' cohorts are relatively large, the loss of precision by deleting values from the alternate survey mode may not be material – this can be examined by comparing standard errors of naïve estimates (i.e. ones which do not attempt to account for mode effects) and those obtained using MI. Nevertheless, the extent of bias arising from violations of the MAR assumption may be considerable and ultimately cannot be determined from the observed data alone.

There is one multiple imputation approach that suffers from neither of these limitations, however: calibration (Jongsma et al., 2023). Calibration uses an external sample who have responded to an item (or set of items) in both survey modes to determine the relationship between responses in each mode. These data are then used as the basis of an imputation model for individuals in the main sample who used the alternate survey mode. Auxiliary information can be incorporated to improve the prediction.

While this method can provide unbiased and more efficient predictions of counterfactual responses for those in the alternate mode, it relies on the same measures being used in the main survey and the calibration sample. It further assumes that mode effects generalize from the calibration sample to the main sample, which may be inappropriate. Currently, there is little calibration data available.

6.4 Sensitivity Analysis

A final approach that has multiple advantages over the previous methods is sensitivity analysis. In sensitivity analysis, the size of the mode effect is assumed and is used to simulate a counterfactual response for those in the alternate survey mode. Substantive models are then run using this simulated data and compared against substantive models using observed values to examine whether, and to what extent, results change. For example, sum scores from a battery of mental health questions may be reduced by, say, 0.3 SD for participants in a telephone mode (which used an interviewer) to obtain counterfactual values for a web survey mode (which was anonymous), with these values then used in analyses, such as regression models (see Section 8 for worked examples). This process can be repeated using a range of values for the mode effect and uncertainty can be incorporated by drawing mode effects from a distribution characterising one's prior beliefs. By correcting data from the alternate mode, sensitivity analysis uses all available information, unlike multiple imputation.

Choosing mode effects can be approached in two ways. The first involves attempting to obtain a plausible value – or set of values – for the mode effect. This may be based upon a search of the prior literature or based upon one's own statistical modelling. For instance, based on general findings that social desirability is a main driver of mode effects in measures of mental health and that effects generally do not exceed 0.3 SD (Section 5). Alternatively, one could estimate the complier average causal effect for a relevant item in the NCDS Sweep 9 mode experiment (Goodman et al., 2022) and apply an (hypothesized) correction factor to simulate any heterogeneity in the mode effects such as for non-compliers or participants who were ineligible for the experiment.

The second approach involves using a fine net of sensitivity analysis parameters to determine the level of mode effects required to materially affect results; if extreme values for mode effects do not change substantive conclusions, results are likely

robust. This approach is most useful in research where only high-level conclusions are being drawn and high-accuracy estimates are not required – for instance, where the goal is determining overall direction of association, rather than estimate parameters for decision models. However, this approach is limited where substantive conclusions do in fact change within the net used for the mode effects.

An advantage of sensitivity analysis is that it is very flexible and can incorporate any complexity into modelling of the mode effect. For instance, heterogeneity can be incorporated by positing unique mode effects for different subsets of participants. This may include different mode effects according to participants characteristics (e.g., male or female) or according to the observed value of the variable to be corrected itself (e.g., larger mode effects for individuals reporting an extreme of a scale than at its mid-point). Mode effects for non-continuous variables can also be handled relatively straightforwardly. For instance, transition matrices can be used to determine the rate of false negative and false positives for binary variables.

Sensitivity analysis has several other advantages. These include a need to only model the mode effect – in studies without mode experiments, such as Next Steps, attempts to estimate mode effects or use multiple imputation appropriately require understanding mode selection in detail. Sensitivity analysis can also be used to correct for mixing modes between sweeps, especially where these modes do not overlap (e.g., making telephone responses commensurate with face-to-face). This is particularly important where an analyst wants to obtain estimates of longitudinal change. The approaches in Sections 6.1-6.3 (excluding calibration) cannot be used in this situation, i.e., where the counterfactual mode is not used by anyone.

Sensitivity analysis can also be combined with the previous approaches. For instance, factors can be posited to correct estimated mode effects to account for mode selection and imputed data can be altered to handle violations of the missingness at random assumption; this approach is known as pattern mixture modelling (Leurent et al., 2018). Sensitivity analysis for mode effects can also be combined with sensitivity analysis carried out for other reasons, for instance to assess whether unobserved confounding may explain results (Rosenbaum, 2019).

A final advantage of sensitivity analysis is transparency. Compared with a statement in a paper’s discussion noting that mode effects may bias results, with sensitivity

analysis, the bias (corresponding to specific assumptions about the mode effects) can be quantified. Analysts can state the size of the mode effect that would be required to materially impact results (Gallop & Weschle, 2019) and discuss whether such a mode effect is plausible or not with reference to the available literature on mode effects. This can focus disagreement and make disagreement more productive – critical reviewers can suggest different parameters to be used in the sensitivity analysis.

Sensitivity analysis is not without disadvantages, however. First, it relies on researcher judgement to determine plausible (or implausible) mode effects. Previous evidence on mode effects may not transport to the current situation which uses a different sample and settings, including survey items that presented slightly differently (e.g., by interviewers with different training) or are different entirely (e.g., if another measure of mental health has been used). Nevertheless, uncertainty can be captured by testing a range of mode effects. Further, where mode selection is expected to operate in the same direction as the mode effect (e.g., both yielding higher values among individuals in the alternate mode), the observed mode difference represents an upper bound on the size of the mode effect. Moreover, the methods in Section 6.1-6.3 also require researcher judgement, particularly in deciding if mode selection has been appropriately modelled, which may be out of a researcher's area of expertise.

Second, to our knowledge, there is no out-of-the-box functionality for performing sensitivity analysis and some manual coding will be needed, requiring programming skills. However, sensitivity analysis is possible in major programming languages and several commands are available that make the procedure relatively straightforward (e.g., in R, `ifelse()` and `case_when()` functions to subtract mode effects from relevant observations and `rnorm()` and related functions that take draws from defined distributions). There are also a number of clear walkthroughs that go through the sensitivity analysis procedure, though focusing on measurement error in general, rather than mode effects (e.g., Gallop & Weschle, 2019; Pina-Sánchez et al., 2023).

Third, where the sensitivity analysis involves multiple parameters – e.g., mode effects for several variables or in multiple modes, or where there are heterogeneous effects across several subgroups – the number of models to run can explode. Condensing and interpreting this information can be difficult. However, methods and software are available for efficiently conveying the results of many models – for instance, plotting

heat maps where there are two parameters and specification curves where there are more (Masur & Scharkow, 2020; Simonsohn et al., 2020; Steegen et al., 2016; Wright, 2023). These are particularly simple to implement using the R package `ggplot2` (Wickham, 2016).

6.5 Coda

We have discussed four approaches to handling mode effects. The first three (statistical control, estimating mode effects, and multiple imputation) are not applicable in all situations and, given that participants ultimately select into mode, it may not be obvious whether the methods are inapplicable in a particular situation or not. Sensitivity analysis is a robust method that can handle uncertainty in researcher judgement and does not require detailed knowledge about mode selection. In the next section, we provide recommendations on handling mode effects in analyses of CLS data.

7 Recommendations for Accounting for Mode Effects

Below we list nine recommendations for handling mode effects. These apply in analyses where data were collected using multiple modes, either within sweep (i.e., where a mixed mode design was used) or between sweeps (e.g., where a telephone survey was preceded by face-to-face sweeps).

First Steps

1. Investigate whether a variable is likely to suffer mode effects, a priori.
    - Section 0 outlines the characteristics of items susceptible to mode effects. d'Ardenne et al. (2017) provide a checklist for scoring items on their likelihood of exhibiting mode effects. If the variable is unlikely to suffer mode effects, disregard Recommendations 2-8, unless it is possible to test this directly (e.g., using data from the NCDS Sweep 9 mode experiment).
2. Determine the likely size of the mode effect based on previous literature.
    - Several experimental studies have been carried out to determine the size and nature of mode effects (see Section 5 for examples). These span multiple samples and characteristics measured (e.g., sociodemographic, attitudes and physical and mental health). To our knowledge, for continuous variables, these almost always do not exceed 0.3 SD, though a judgment about the transportability of results needs to be made.
3. Draw out your assumptions of the mode effect and mode selection processes relevant to your substantive analysis using DAGs.
    - Determine from the DAG whether it is possible to unbiasedly estimate mode effects.

Analysis

4. Report descriptive statistics on survey mode, including the proportions of the sample in each mode and the characteristics of participants in each.

- For the descriptive statistics, use measures that are unlikely to suffer from mode effects, if possible. Otherwise, mode differences may reflect mode effects rather than mode selection. These descriptive statistics should be additional to the standard Table 1 descriptive statistics produced for the analysis.

5. Run a ‘naïve’ analysis not accounting for survey mode.
6. (If supported by the DAG), run the substantive analysis accounting for mode effects.
    - Use either the (a) statistical control, (b) estimating the counterfactual value directly, or (c) multiple imputation approaches.
    - The loss of statistical power from using the multiple imputation approach can be gauged by comparing standard errors from the naïve and MI analyses and by examining the number of participants in each survey mode.
7. Run a sensitivity analysis positing values for the mode effects.
    - Assume plausible or implausible values (ideally both) and examine how results change according to mode effects assumed. Perform an analysis even if mode effects can in principle be estimated unbiasedly as unbiased estimates may only be obtainable in practice in a subsample of participants (e.g., those eligible in a mode experiment) and may not transport to other groups. Reviewers and other researchers may also disagree with the arguments supporting the belief that mode effects can be estimated unbiasedly given available data.

Reporting

8. Report the results of sensitivity analysis and (if applicable) other analyses performed accounting for mode effects.
    - Describe whether results change quantitatively over the range of mode effect parameters examined. If possible, state what level of mode effect would be required to change results qualitatively.
9. Discuss mode effects in strengths and limitations sections.
    - Describe transparently the likelihood that mode effects are (a) present and (b) could alter substantive conclusions.

8 Worked Examples

To demonstrate the approaches outlined in Section 6, in this section we show worked examples using mixed mode data from Sweep 9 (55y) of the NCDS and Sweep 6 (18/19y) of Next Steps. For illustration, we focus our examples on linear regression. Walkthroughs including annotated R and Stata code are available at https://osf.io/kq5ra.

8.1 The Association Between Social Class and Quality of Life: NCDS Sweep 9 (55y)

A large literature has found evidence of pervasive social gradients in health, such that those in more advantaged socioeconomic positions have better health on average (Marmot, 2015, 2016). This is true of almost all societies, times, and dimensions of health, including measures of mental health and wellbeing (Allen et al., 2014; Olsen et al., 2020). In this example, we examine the cross-sectional association between social class (professional vs not) and quality of life as measured with the CASP-6 Likert Score (Hyde et al., 2003; Wiggins et al., 2017) using data from Sweep 9 (55y) of the NCDS.

As Goodman et al. (2022) show, items of the CASP-6 in the NCDS mode experiment exhibit sizeable mode effects – participants who responded by telephone reported ~ 0.25 SD higher wellbeing, on average. While there was no evidence that reporting professional social class exhibited mode effects, professional social class at age 50y was strongly related to web participation in the age 55y survey. We would therefore expect a naïve analysis not accounting for mode effects to understate the association between professional social class and the CASP-6. A DAG representing our hypothesis is shown in Figure 8.1. The estimate from a simple regression of CASP-6 scores on professional social class is shown in Figure 8.2; individuals in professional social class report 1.20 (95% CI = 1.05, 1.36) points higher CASP-6 Likert scores on average, indicating greater wellbeing.

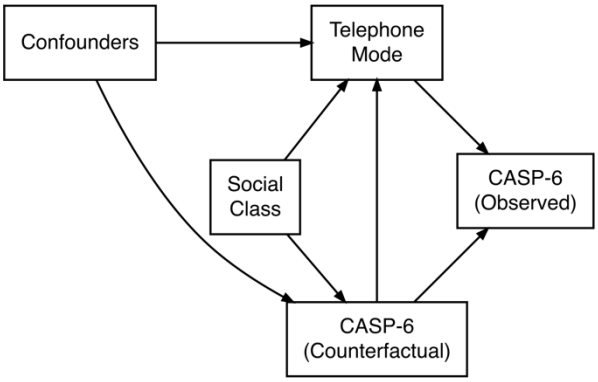

Figure 8.1: DAG representing assumed mode effect and mode selection processes relevant to examining differences in quality of life (CASP-6 Likert scores) according to professional social class in Sweep 9 (55y) of the NCDS. Low quality of life and non-professional social class are hypothesised to be causes of selection into telephone survey mode. CASP-6 scores are hypothesised to exhibit mode effects, such that participants in (non-anonymous) telephone mode report systematically higher wellbeing. There are also a set of unmeasured variables that are causes of mode selection and determinants of CASP-6 scores.

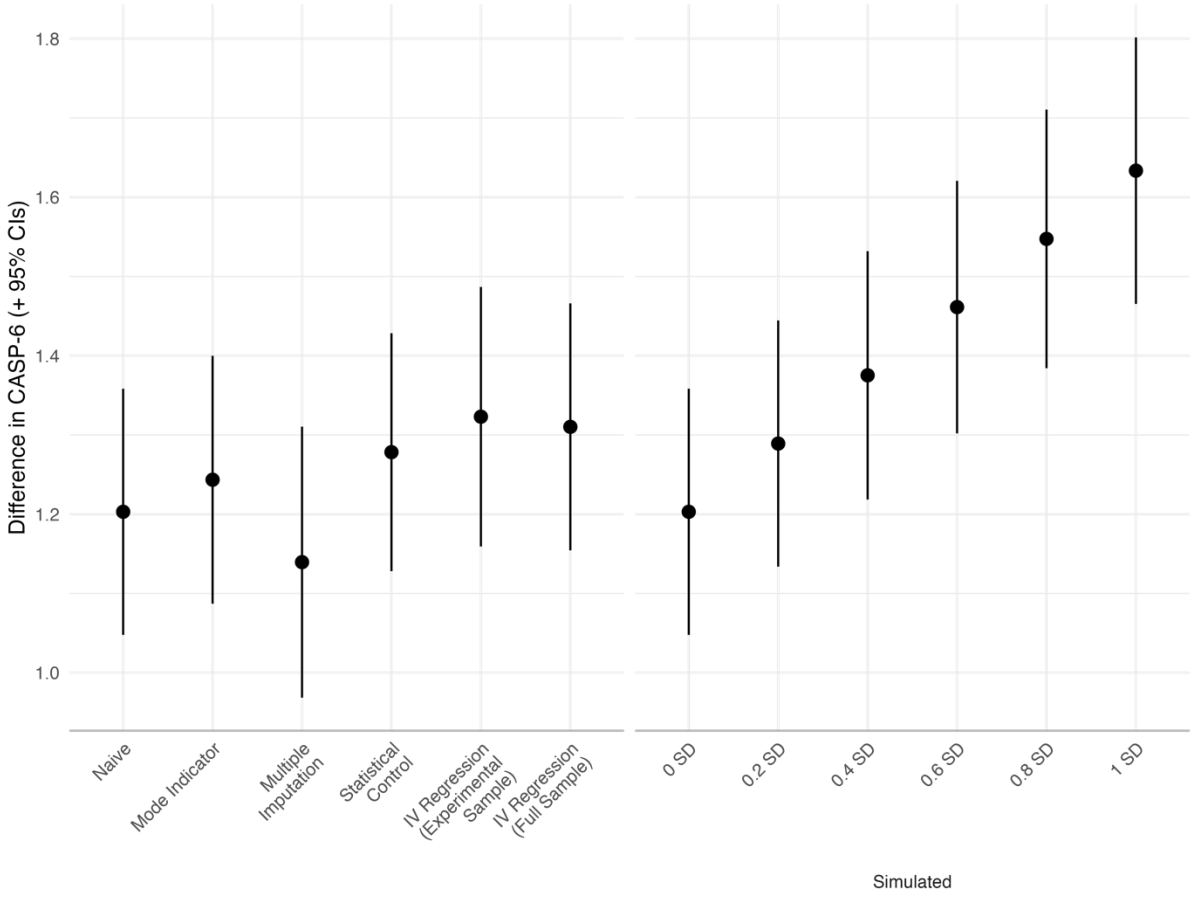

Figure 8.2: Inequalities in quality of life by social class, age 55y of the NCDS survey. Estimates from regression models with different approaches used to attempt to account for mode effects in the quality of life (CASP-6) variable.

Including an indicator variable for survey (telephone) mode increases the size of the association by approximately 3% (1.24 points, 95% CI = 1.09, 1.40; Figure 8.2). However, this likely still yields a biased association. Social class is only one predictor of mode selection and not accounting for others could generate collider bias (Figure 8.1). Further, quality of life at age 50y is a predictor of mode selection: web participants in the mixed mode experimental arm had greater wellbeing, on average (Goodman et al., 2022). This plausibly proxies for contemporaneous quality of life, suggesting age 55y CASP-6 scores predict mode selection too. Adjusting for mode may therefore attenuate associations due to range restriction (Section 4.6). This would also bias estimates obtained using multiple imputation; mode selection according to quality of life would imply that CASP-6 values for the web mode are missing not at random.

Indeed, the MI estimate is attenuated relative to the naïve model (1.14 points, 95% CI = 0.97, 1.31; Figure 8.2).[15]

Next, we examine what happens when using statistical control to try to account for the mode effect. As we are interested in the bivariate association between professional social class and CASP-6 scores, we do not attempt to account for mode selection directly by adding control variables to the main substantive regression model as this would change the interpretation of the estimated quantity. Instead, we use a two-step approach, estimating the mode effect and using this to generate (predicted) counterfactual values. Specifically, we use linear regression to estimate differences in age 55y CASP-6 scores by mode, controlling for sex and three measures related to mode selection (Goodman et al., 2022)[16], and then subtract the estimated mode effect from observed CASP-6 scores for all participants in the telephone mode (eligible for the mode experiment or not) in order to obtain a (predicted) counterfactual value for the web survey mode. Next, we regress these values on professional social class, calculating confidence intervals with bootstrapping (1000 samples) to propagate uncertainty in the mode effect estimate. This yields an association between professional social class and the CASP-6 measure of 1.28 points (95% CI = 1.13, 1.43; Figure 8.2), 6% larger than the naïve estimate.

However, the procedure may still yield a biased estimate – the control variables used may not fully capture mode selection. To control for survey mode appropriately, we next use instrumental variable regression with survey mode instrumented by experimental arm. This yields a slightly stronger association between professional social class and the CASP-6 measure of 1.32 points (95% CI = 1.16, 1.49; Figure 8.2). One issue with this analysis, however, is that not all participants were eligible for the mixed mode experiment; this single instrumental variable regression approach discards their data. To get around this, we use a similar two-step approach to that used for the statistical control method above: we first estimate the mode effect using instrumental variables regression and then subtract this from observed CASP-6 scores for all participants in the telephone mode before regressing these values on professional social class, calculating confidence intervals with bootstrapping (1000

[15] We included only a few auxiliary variables in our imputation model: sex, CASP-12 score at age 50y, and two measures of cognitive ability at age 50y. Full analyses would likely select a more complete set of auxiliary variables.
[16] CASP-12 scores and two measures of cognitive ability, each measured at age 50y.

samples) to propagate uncertainty in the mode effect estimate. The results are shown in Figure 8.2 and are similar to those from the IV analysis discarding ineligible participants: a 1.31 point (95% CI = 1.15, 1.47; Figure 8.2) differences in CASP-6 scores according to social class.[17] Note, the mode effect is sizable at approximately 0.9 CASP-6 Likert points (~ 0.25 SD). This is comparable to the association between current unemployment and mental distress found in longitudinal studies (Paul & Moser, 2009).[18]

As not all individuals assigned to the sequential web experimental arm participated via web, the instrumental variable analysis recovers the 'complier average causal effect' (CACE) among those who answered via web (~ 75% of the mixed mode experimental sample). It assumes that those who responded by telephone did not change their answers due to being offered web. Even though we can plausibly estimate the CACE, it is still worth using sensitivity analysis to examine mode effects as the CACE assumption might not hold in practice or may not reflect the average mode effect in the full sample (i.e., compliers, non-compliers and participants ineligible for the experiment). In Figure 8.2, we show the range of estimates assuming constant mode effects between 0 and 1 SD (approximately equivalent to 0-3.5 CASP-6 Likert points). Mode effects over 0.5 SD are implausible, given previous literature (Section 5), but a mode effect of 1 SD increases the association by only 36% and the substantive conclusion is the same: individuals in the professional social class have higher wellbeing.[19] In our opinion, determining this is quantitatively is better than a vague line added to a limitations section on the potential, in principle, for mode effects to impact results. It is also preferable to solely relying on a method that could be biased by mode selection.

[17] As discussed in Section 5, experimental arm also had effects on unit and item non-response. Unaccounted for, this can induce selection biases. Here, we use complete case data and added no further control variables to focus on the central material. In practice, more principled approaches to missingness should be used. For instance, Goodman et al. (2022) handle missingness with multiple imputation.

[18] The estimated mode effect using the statistical control approach (which is likely biased due to residual mode selection) was smaller: 0.63 CASP-6 Likert points (~ 0.17 SD)

[19] For context, the difference in male and female height is approximately 1.8 SD (calculated from figures in Roser et al., 2023)

8.2 Gender Differences in Adolescent Sexual Initiation: Next Steps Sweep 6 (18/19y)

Accurately measuring sexual activity is challenging (Dare & Cleland, 1994; Fenton et al., 2001). Measurement largely relies upon self-report and depending on cultural norms and expectations, survey respondents may over- or under-report sexual activity, particularly in the presence of an interviewer (Burkill et al., 2016; Copas et al., 2002; Fenton et al., 2001; Wadsworth et al., 1993). Further, there tend to be gender differences in inaccurate reports: men report more heterosexual partners than women, even though in closed populations these numbers should be equal (Curtis & Sutherland, 2004; Nnko et al., 2004; Wadsworth et al., 1993). This discrepancy is thought to be due to both over-reporting by men and under-reporting by women, reflecting different social expectations placed on these groups ('secretive females' and 'swaggering males', Nnko et al., 2004; see also, Dare & Cleland, 1994; Fenton et al., 2001). These reporting biases may influence estimates derived from survey responses, particularly in mixed mode settings. In this example, we examine gender differences in reporting ever having sexual intercourse as measured in Sweep 6 (18/19y) of Next Steps.

Sweep 6 of Next Steps used a sequential mixed mode design with web, telephone and face-to-face interview offered in turn. Questions on sexual activity were included in each mode, though differed in their presentation. In the web survey, no interviewer was present, while in the face-to-face interview, questions were asked in a self-completion module with participants handed a computer to read and answer the questions themselves. In the telephone survey, interviewers read the questions aloud and participants answered verbally with yes or no answers. Telephone responses were therefore not anonymous and could feasibly have been influenced by social desirability concerns. These responses may therefore exhibit mode effects relative to face-to-face and web modes. Respondents in the telephone survey (48% of the sample) were also disproportionately male. Combined, this is likely to bias estimates of gender differences in reported sexual activity. Specifically, assuming male telephone respondents are more likely overstate ever having sex compared with females (who may understate instead), we would anticipate the association between sex and sexual intercourse to be biased towards appearing that males are more likely

to have had sex than would have occurred if only face-to-face or web survey modes were used.[20]

Before exploring the association between gender and sexual intercourse, to show the difficulty of estimating mode effects in Next Steps, we estimate the association between number of children at age 25 and survey mode at age 25 – we assume this should be subject to little mode effect if any, so any association should reflect selection rather than a mode effect (i.e., we use it as a negative control outcome; Lipsitch et al., 2010). The unadjusted association shows differences in the probability of having children, relative to web survey (which was offered first), of -2.32 pp. (95% CI = -5.47 pp., 0.82 pp.) in the telephone survey and 14.73 pp. (95% CI = 12.75 pp., 16.71 pp.) in the face-to-face survey. Adjusting for a large suite of controls, the latter association weakens but only to 8.44 pp. (95% CI = 5.83 pp., 11.04 pp.), a sizeable difference.[21]

As in the previous example, we first naively estimate a univariate regression of ever having sex upon gender that does not include mode as a covariate (Figure 8.3). The difference is -3.99 pp. (95% CI = -5.77 pp., -2.22 pp.), with females less likely to report ever having had sex. Next, we include survey mode as a covariate; the association decreases to -2.44 pp. (95% CI = -4.23 pp., -0.65 pp.; Figure 8.3). However, this is unlikely to have eliminated bias – gender is only one predictor of mode selection. Using multiple imputation, the association reverses sign: 1.46 pp. (95% CI = -0.72 pp., 3.64 pp.; Figure 8.3). Note, given multiple imputation discards data in this setting, the association is less precisely estimated – the size of the standard error is increased by 22%.

[20] In this example, for brevity, we assume that only telephone responses exhibit mode effects (relative to web and face-to-face modes). In practice, the face-to-face mode is likely to exhibit mode effects (relative to the web survey) even though responses were given as part of a self-completion module. In their analysis of items on sexual activity given in a self-complete module of a face-to-face interview and later asked again via a (fully anonymous) web survey, Burkill et al. (2016) find the physical presence of an interviewer alone is sufficient to induce mode effects.

[21] The controls were sex, maternal and paternal age at birth, family type (0, 1, or 2 parent household), family socioeconomic class, highest parental education, cohort member's activity, whether they received educational maintenance allowance, self-rated health, long-standing illness, ever used cannabis, alcohol consumption, special educational needs, and whether the cohort member had ever been in care.

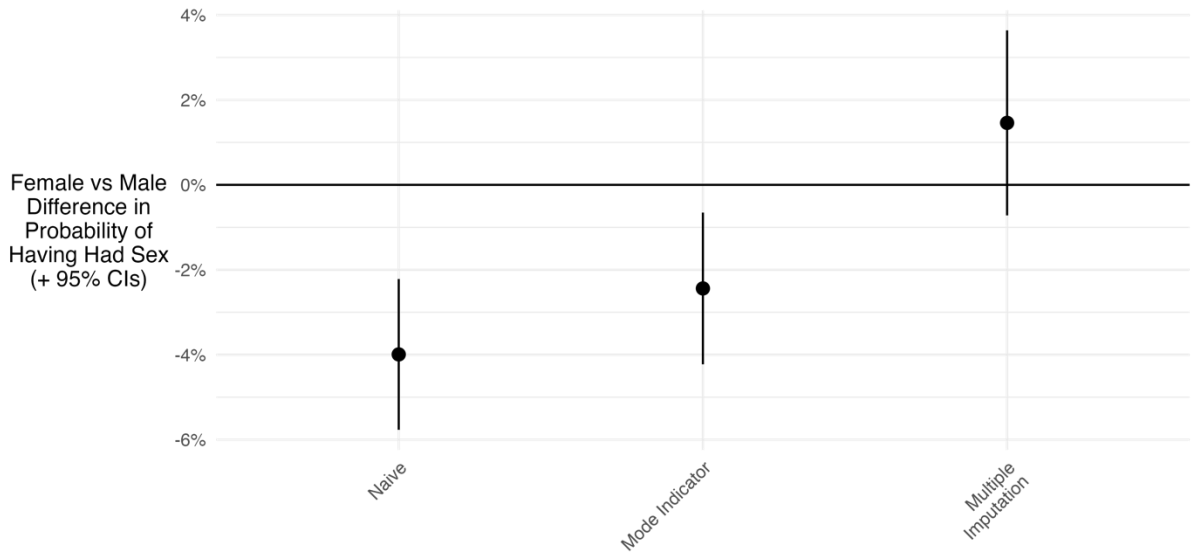

Figure 8.3: Gender differences in the probability of ever having had sex, age 18/19y of the Next Steps survey. Estimates from regression models with different approaches used to attempt to account for mode effects in reports of ever having had sex for respondents in the telephone survey mode.

We are unlikely to be able to estimate the mode effect – we do not have good theories of why people select into mode, so it is not sensible to try to estimate the mode effect using statistical adjustment, and, as far as we are aware, there is no source of random variation in survey mode used in Next Steps that could be exploited as an instrumental variable. Instead, we can perform sensitive analysis, simulating the mode effect using external information. As noted, evidence from the sex research literature suggests men typically overstate, and females understate, sexual activity. For binary variables (e.g., ever had sex vs never had sex), there are two forms of misclassification – false negatives (individuals reporting having not had sex who in fact had) and false positives (individuals reporting having had sex who in fact had not). In our simulations, we assume that the false negative and false positives rates are between 0% (i.e., no measurement error) and 10%, selecting values in 2% increments (i.e., we assume false negative and false positive rates of 0%, 2%, 4%, 6%, 8%, and 10%). This is arguably conservative. While it is generally not possible to validate self-reports of sexual activity, inconsistencies between successive reports can be instructive: in a US sample, Upchurch et al. (2002) find 5.9% of White girls and 12.3% of White boys 'reclaim' virginity in a later sweep of a survey.

Next, we constrain our sensitivity analysis by incorporating four beliefs:

1. For males, false positives should exceed false negatives.

2. For females, false negatives should exceed false positives.
3. The false positive rate for males should exceed the false positive rate for females.
4. The false negative rate for females should exceed the false negative rate for males.

Otherwise, we test each combination of male and female false positive and false negative rates (301 models). Reasoning like this reduces the simulation space substantially.

Figure 8.4 shows the result of the sensitivity analysis; estimates are ordered on the x-axis according to rank (this is known as a specification curve; Simonsohn et al., 2020). The figure shows that there is substantial uncertainty in the estimates both quantitatively and substantively. Depending on the parameters assumed, the sign of the association reverses direction and in approximately two thirds of simulations is statistically insignificant. The estimate at the 75th percentile is just a -0.21% (95% CI = -2.05%, 1.64%) difference and corresponds to false positive rate of 10% and 6% and false negative rates of 2% and 8% among males and females, respectively.

Assuming the range of parameters chosen is sensible, the naïve estimate represents close to a lower bound on the association and may incorrectly give the appearance of females being less likely to have had sex. If we had assumed stronger mode effects, the variability of the estimates would have been greater still. Again, we believe that being able to show this transparently is preferable to adding a nebulous and non-committal line on mode effects to a limitations section.

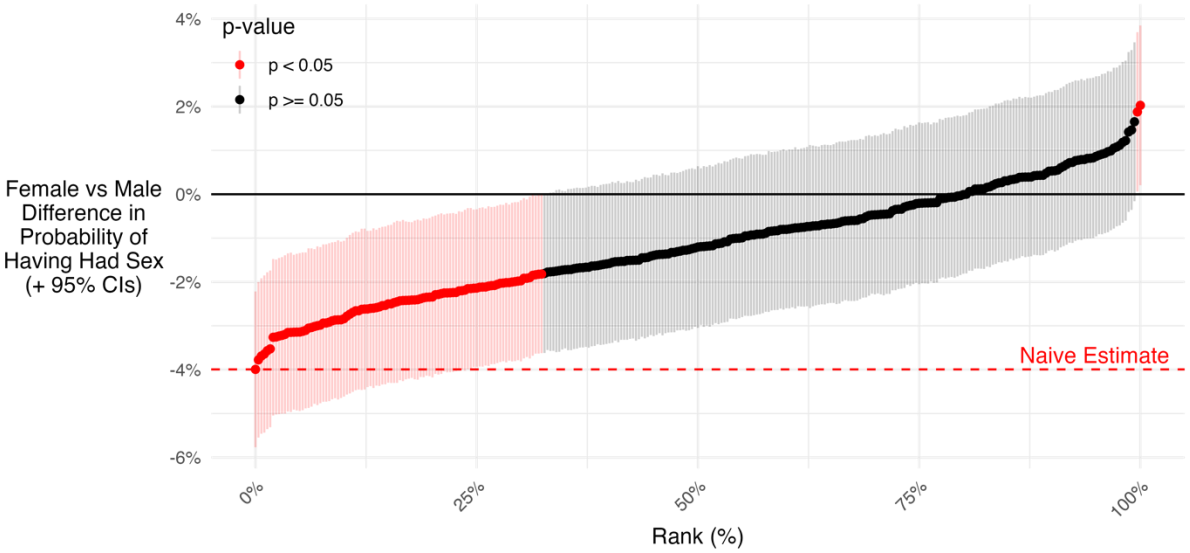

Figure 8.4: Gender differences in the probability of ever having had sex, age 18/19y of the Next Steps survey. Estimates from sensitivity analysis assuming different likelihoods of false positives and false negative reports in the telephone survey by gender.

The specification curve in Figure 8.4 efficiently conveys the main results of the sensitive analysis. However, it does not show which parameters or combination of parameters drive results. Heat maps can be effective, though are limited when there are more than two parameters in the sensitivity analysis. An alternative approach is to use mixed effects modelling to calculate the proportion of variance in the sensitivity analysis estimates that can be explained by each of the parameters included (Masur & Scharkow, 2020). This can help identify which parameters the results are particularly sensitive to (at least over the range used in the sensitivity analysis). Figure 8.5 shows that results were most sensitive to the male false positive probability parameter (responsible for 64% of the total variance), which makes sense given that most participants reported ever having had sex and males were more likely to participate using the telephone mode. The results in Figure 8.5 suggest that obtaining more precise values for the false positive rates would pay the greatest dividend in providing more accurate mode effect adjusted results.

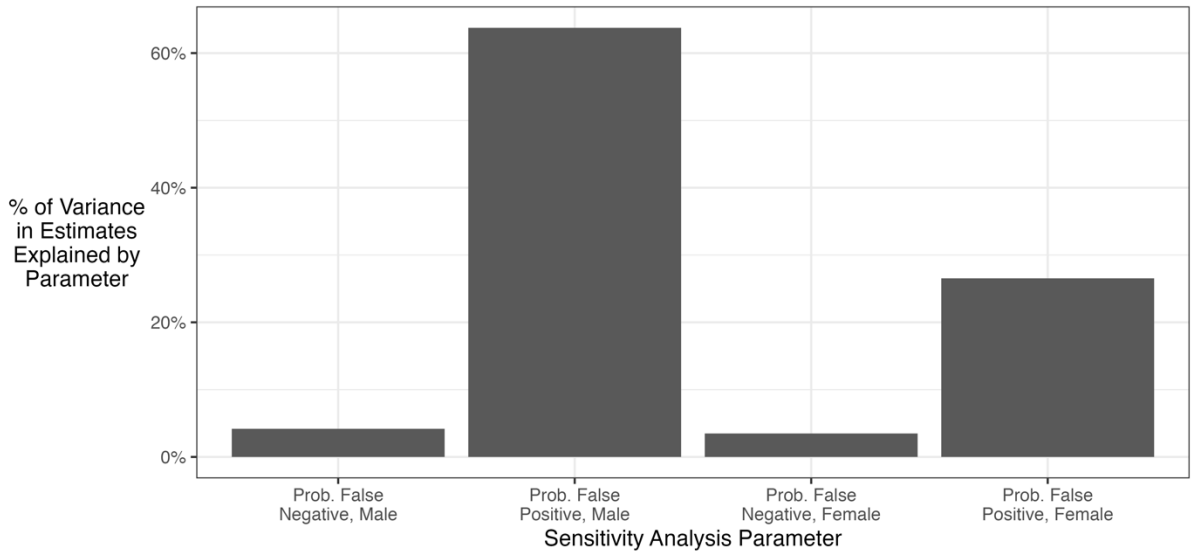

Figure 8.5: Variance decomposition of the estimates obtained from the sensitivity analysis of Next Steps data accounting for mode effects in reports of sexual activity. Derived from linear mixed effects model (Masur & Scharkow, 2020).

8.3 Discussion

In this section, we have seen two worked examples accounting for mode effects. The sensitivity analyses in each case modelled the mode effects relatively straightforwardly. In real-world analyses, more complexity may be warranted – for instance, using a larger set of participant characteristics to model heterogeneity in the mode effect or incorporating uncertainty in beliefs about the mode effects by specifying a distribution the mode effect is drawn from. Nevertheless, these relatively straightforward analyses yielded important information. In the first example, implausibly large mode effects did not change substantive conclusions, while in the second example, estimates were very sensitive, in both size and sign, to the possibility of false reports in the telephone mode. These statements are transparent and to our minds more useful than lines such as ‘results may be biased by mode effects’ added to discussion sections. Supported by empirical evidence, they focus attention toward or away from mode effects as important factors that may explain results.

9 References

Al Baghal, T. (2019). The Effect of Online and Mixed-Mode Measurement of Cognitive Ability. *Social Science Computer Review*, *37*(1), 89–103. https://doi.org/10.1177/0894439317746328

Allen, J., Balfour, R., Bell, R., & Marmot, M. (2014). Social determinants of mental health. *International Review of Psychiatry*, *26*(4), 392–407. https://doi.org/10.3109/09540261.2014.928270

Bann, D., Wright, L., Hughes, A., & Chaturvedi, N. (2023). Socioeconomic inequalities in cardiovascular disease: A causal perspective. *Nature Reviews Cardiology*, 1–12. https://doi.org/10.1038/s41569-023-00941-8

Bell, A. (2014). Life-course and cohort trajectories of mental health in the UK, 1991-2008—A multilevel age-period-cohort analysis. *Social Science and Medicine*, *120*, 21–30. https://doi.org/10.1016/j.socscimed.2014.09.008

Boniface, S., Scholes, S., Shelton, N., & Connor, J. (2017). Assessment of Non-Response Bias in Estimates of Alcohol Consumption: Applying the Continuum of Resistance Model in a General Population Survey in England. *PLOS ONE*, *12*(1), e0170892. https://doi.org/10.1371/journal.pone.0170892

Brown, M. (2016). *Going online with the National Child Development Study: Design decisions during the development of the Age 55 web survey* (Working Paper 2016/2; p. 24). Centre for Longitudinal Studies. https://cls.ucl.ac.uk/wp-content/uploads/2017/04/CLS-WP-20162.pdf

Brown, M., & Calderwood, L. (2020). *Mixing modes in longitudinal surveys: An overview* (2020/3; CLS Working Paper, p. 11). UCL Centre for Longitudinal Studies. https://cls.ucl.ac.uk/wp-content/uploads/2020/04/CLS-working-paper-2020-3-Mixing-modes-in-longitudinal-surveys-an-overview.pdf

Buelens, B., & Brakel, J. A. V. den. (2017). Comparing Two Inferential Approaches to Handling Measurement Error in Mixed-Mode Surveys. *Journal of Official Statistics*, *33*(2), 513–531. https://doi.org/10.1515/jos-2017-0024

Burkill, S., Copas, A., Couper, M. P., Clifton, S., Prah, P., Datta, J., Conrad, F., Wellings, K., Johnson, A. M., & Erens, B. (2016). Using the Web to Collect Data on Sensitive Behaviours: A Study Looking at Mode Effects on the British National Survey of Sexual Attitudes and Lifestyles. *PLOS ONE*, *11*(2), e0147983. https://doi.org/10.1371/journal.pone.0147983

Calderwood, L., Peycheva, D., Henderson, M., Mostafa, T., & Rihal, S. (2017). *Next Steps: Sweep 8-Age 25 Survey User Guide (First Edition)* (p. 37). http://doi.org/10.5255/UKDA-SN-5545-6

Cernat, A. (2015). Impact of mixed modes on measurement errors and estimates of change in panel data. *Survey Research Methods*, *9*(2), Article 2. https://doi.org/10.18148/srm/2015.v9i2.5851

Cernat, A., Couper, M. P., & Ofstedal, M. B. (2016). Estimation of Mode Effects in the Health and Retirement Study Using Measurement Models. *Journal of Survey Statistics and Methodology*, *4*(4), 501–524. https://doi.org/10.1093/jssam/smw021

Cernat, A., & Sakshaug, J. W. (2021). Estimating the Measurement Effects of Mixed Modes in Longitudinal Studies: Current Practice and Issues. In *Advances in Longitudinal Survey Methodology* (pp. 227–249). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119376965.ch10

Clarke, P. S., & Bao, Y. (2022). Estimating mode effects from a sequential mixed-mode experiment using structural moment models. *The Annals of Applied Statistics*, *16*(3), 1563–1585. https://doi.org/10.1214/21-AOAS1557

Copas, A. J., Wellings, K., Erens, B., Mercer, C. H., McManus, S., Fenton, K. A., Korovessis, C., Macdowall, W., Nanchahal, K., & Johnson, A. M. (2002). The accuracy of reported sensitive sexual behaviour in Britain: Exploring the extent of change 1990-2000. *Sexually Transmitted Infections*, *78*(1), 26. https://doi.org/10.1136/sti.78.1.26

Curtis, S., & Sutherland, E. (2004). Measuring sexual behaviour in the era of HIV/AIDS: The experience of Demographic and Health Surveys and similar enquiries. *Sexually Transmitted Infections*, *80*(Suppl 2), ii22–ii27. https://doi.org/10.1136/sti.2004.011650

d'Ardenne, J., Collins, D., Gray, M., Jessop, C., & Pilley, S. (2017). *Assessing the risk of mode effects: Review of proposed survey questions for waves 7-10 of Understanding Society* (Working Paper 2017–04; Understanding Society Working Paper Series, p. 18). University of Essex. https://www.understandingsociety.ac.uk/research/publications/524254

Dare, O. O., & Cleland, J. G. (1994). Reliability and validity of survey data on sexual behaviour. *Health Transition Review*, *4*, 93–110.

Department for Education. (2011). *LSYPE User Guide to the Datasets: Wave 1 to Wave 7* (p. 103). http://doi.org/10.5255/UKDA-SN-5545-6

Digitale, J. C., Martin, J. N., & Glymour, M. M. (2022). Tutorial on directed acyclic graphs. *Journal of Clinical Epidemiology*, *142*, 264–267. https://doi.org/10.1016/j.jclinepi.2021.08.001

Domingue, B. W., McCammon, R. J., West, B. T., Langa, K. M., Weir, D. R., & Faul, J. (2023). The Mode Effect of Web-Based Surveying on the 2018 U.S. Health and Retirement Study Measure of Cognitive Functioning. *The Journals of*

Gerontology: Series B, *78*(9), 1466–1473. https://doi.org/10.1093/geronb/gbad068

Elwert, F., & Winship, C. (2014). Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable. *Annual Review of Sociology*, *40*(1), 31–53. https://doi.org/10.1146/annurev-soc-071913-043455

Fenton, K. A., Johnson, A. M., McManus, S., & Erens, B. (2001). Measuring sexual behaviour: Methodological challenges in survey research. *Sexually Transmitted Infections*, *77*(2), 84–92. https://doi.org/10.1136/sti.77.2.84

Gallop, M., & Weschle, S. (2019). Assessing the Impact of Non-Random Measurement Error on Inference: A Sensitivity Analysis Approach. *Political Science Research and Methods*, *7*(2), 367–384. https://doi.org/10.1017/psrm.2016.53

Gondek, D., Bann, D., Patalay, P., Goodman, A., McElroy, E., Richards, M., & Ploubidis, G. B. (2021). Psychological distress from early adulthood to early old age: Evidence from the 1946, 1958 and 1970 British birth cohorts. *Psychological Medicine*, 1–10. https://doi.org/10.1017/S003329172000327X

Goodman, A., Brown, M., Silverwood, R. J., Sakshaug, J. W., Calderwood, L., Williams, J., & Ploubidis, G. B. (2022). The Impact of Using the Web in a Mixed-Mode Follow-up of a Longitudinal Birth Cohort Study: Evidence from the National Child Development Study. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *185*(3), 822–850. https://doi.org/10.1111/rssa.12786

Hernán, M. A. (2018). *HarvardX: Causal Diagrams: Draw Your Assumptions Before Your Conclusions*. edX. https://www.edx.org/learn/data-analysis/harvard-university-causal-diagrams-draw-your-assumptions-before-your-conclusions

Hernan, M. A., & Robins, J. M. (2023). *Causal inference: What if* (First edition). Taylor and Francis.

Hox, J., De Leeuw, E., & Klausch, T. (2017). Mixed-Mode Research: Issues in Design and Analysis. In P. P. Biemer, E. Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, N. C. Tucker, & B. T. West (Eds.), *Total Survey Error in Practice* (1st ed., pp. 511–530). Wiley. https://doi.org/10.1002/9781119041702.ch23

Hox, J. J., De Leeuw, E. D., & Zijlmans, E. A. O. (2015). Measurement equivalence in mixed mode surveys. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00087

Hutcheon, J. A., Chiolero, A., & Hanley, J. A. (2010). Random measurement error and regression dilution bias. *BMJ*, *340*, c2289. https://doi.org/10.1136/bmj.c2289

Hyde, M., Wiggins, R. D., Higgs, P., & Blane, D. B. (2003). A measure of quality of life in early old age: The theory, development and properties of a needs satisfaction model (CASP-19). *Aging & Mental Health*, *7*(3), 186–194. https://doi.org/10.1080/1360786031000101157

Jäckle, A. (2016). *Identifying and Predicting the Effects of Data Collection Mode on Measurement.* [Workshop Paper].

Jäckle, A., Gaia, A., & Benzeval, M. (2017). *Mixing modes and measurement methods in longitudinal studies* (p. 28) [Resource Report]. CLOSER. https://www.closer.ac.uk/wp-content/uploads/Mixing-modes-and-measurement-methods-in-longitudinal-studies-FULL-FINAL.pdf

Jäckle, A., Roberts, C., & Lynn, P. (2006). *Telephone versus Face-to-Face Interviewing: Mode Effects on Data Quality and Likely Causes* (Working Paper

2006–41; ISER Working Paper, p. 97). Institute for Social and Economic Research.

Jongsma, H. E., Moulton, V. G., Ploubidis, G. B., Gilbert, E., Richards, M., & Patalay, P. (2023). Psychological Distress Across Adulthood: Equating Scales in Three British Birth Cohorts. *Clinical Psychological Science*, *11*(1), 121–133. https://doi.org/10.1177/21677026221095856

Kahneman, D. (2012). *Thinking, fast and slow*. Penguin Books.

Kim, Y., Dykema, J., Stevenson, J., Black, P., & Moberg, D. P. (2019). Straightlining: Overview of Measurement, Comparison of Indicators, and Effects in Mail–Web Mixed-Mode Surveys. *Social Science Computer Review*, *37*(2), 214–233. https://doi.org/10.1177/0894439317752406

Kolenikov, S., & Kennedy, C. (2014). Evaluating Three Approaches to Statistically Adjust for Mode Effects. *Journal of Survey Statistics and Methodology*, *2*(2), 126–158. https://doi.org/10.1093/jssam/smu004

Krosnick, J. A., & Alwin, D. F. (1987). An Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement. *Public Opinion Quarterly*, *51*(2), 201–219.

Lawlor, D. A., Tilling, K., & Smith, G. D. (2016). Triangulation in aetiological epidemiology. *International Journal of Epidemiology*, *45*(6), 1866–1886. https://doi.org/10.1093/ije/dyw314

Leurent, B., Gomes, M., Faria, R., Morris, S., Grieve, R., & Carpenter, J. R. (2018). Sensitivity Analysis for Not-at-Random Missing Data in Trial-Based Cost-Effectiveness Analysis: A Tutorial. *PharmacoEconomics*, *36*(8), 889–901. https://doi.org/10.1007/s40273-018-0650-5

Lipsitch, M., Tchetgen Tchetgen, E., & Cohen, T. (2010). Negative Controls: A Tool for Detecting Confounding and Bias in Observational Studies. *Epidemiology*, *21*(3), 383–388. https://doi.org/10.1097/EDE.0b013e3181d61eeb

Marmot, M. (2015). *Status syndrome: How your social standing directly affects your health* (1st ed.). Bloomsbury.

Marmot, M. (2016). *The health gap: The challenge of an unequal world* (1st ed.). Bloomsbury.

Masur, P. K., & Scharkow, M. (2020). *specr: Statistical functions for conducting specification curve analyses* (Version 0.2.1) [R]. https://CRAN.R-project.org/package=specr

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan* (2nd ed.). Taylor and Francis, CRC Press.

Morgan, S. L., & Winship, C. (2015). *Counterfactuals and causal inference: Methods and principles for social research* (Second edition). Cambridge University press.

Nnko, S., Boerma, J. T., Urassa, M., Mwaluko, G., & Zaba, B. (2004). Secretive females or swaggering males?: An assessment of the quality of sexual partnership reporting in rural Tanzania. *Social Science & Medicine*, *59*(2), 299–310. https://doi.org/10.1016/j.socscimed.2003.10.031

Norris, T., Bann, D., Hardy, R., & Johnson, W. (2020). Socioeconomic inequalities in childhood-to-adulthood BMI tracking in three British birth cohorts. *International Journal of Obesity*, *44*(2), 388–398. https://doi.org/10.1038/s41366-019-0387-z

Ofstedal, M. B., Kézdi, G., & Couper, M. P. (2022). Data quality and response distributions in a mixed-mode survey. *Longitudinal and Life Course Studies*, 1–26. https://doi.org/10.1332/175795921X16494126913909

Olsen, J. A., Lindberg, M. H., & Lamu, A. N. (2020). Health and wellbeing in Norway: Population norms and the social gradient. *Social Science & Medicine*, *259*, 113155. https://doi.org/10.1016/j.socscimed.2020.113155

Øverland, S., Woicik, W., Sikora, L., Whittaker, K., Heli, H., Skjelkvåle, F. S., Sivertsen, B., & Colman, I. (2020). Seasonality and symptoms of depression: A systematic review of the literature. *Epidemiology and Psychiatric Sciences*, *29*, e31. https://doi.org/10.1017/S2045796019000209

Paul, K. I., & Moser, K. (2009). Unemployment impairs mental health: Meta-analyses. *Journal of Vocational Behavior*, *74*(3), 264–282. https://doi.org/10.1016/j.jvb.2009.01.001

Pearl, J., & Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect* (First edition). Basic Books.

Pina-Sánchez, J., Brunton-Smith, I., Buil-Gil, D., & Cernat, A. (2023). Exploring the impact of measurement error in police recorded crime rates through sensitivity analysis. *Crime Science*, *12*(1), 14. https://doi.org/10.1186/s40163-023-00192-5

Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *54*(3), 507–554.

Roberts, C., Gilbert, E., Allum, N., & Eisner, L. (2019). Research Synthesis: Satisficing in Surveys: A Systematic Review of the Literature. *Public Opinion Quarterly*, *83*(3), 598–626. https://doi.org/10.1093/poq/nfz035

Rosenbaum, P. R. (2019). *Observation and experiment: An introduction to causal inference*. Harvard University Press.

Roser, M., Appel, C., & Ritchie, H. (2023). Human Height. *Our World in Data*. https://ourworldindata.org/human-height

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470316696

Rubin, D. B. (2005). Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. *Journal of the American Statistical Association*, *100*(469), 322–331.

Sakshaug, J., Cernat, A., Silverwood, R. J., Calderwood, L., & Ploubidis, G. B. (2022). Measurement Equivalence in Sequential Mixed-Mode Surveys. *Survey Research Methods*, 29-43 Pages. https://doi.org/10.18148/SRM/2022.V16I1.7811

Salganik, M. J., Lundberg, I., Kindel, A. T., Ahearn, C. E., Al-Ghoneim, K., Almaatouq, A., Altschul, D. M., Brand, J. E., Carnegie, N. B., Compton, R. J., Datta, D., Davidson, T., Filippova, A., Gilroy, C., Goode, B. J., Jahani, E., Kashyap, R., Kirchner, A., McKay, S., … McLanahan, S. (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, *117*(15), 8398–8403. https://doi.org/10.1073/pnas.1915006117

Seligman, B., Tuljapurkar, S., & Rehkopf, D. (2018). Machine learning approaches to the social determinants of health in the health and retirement study. *SSM - Population Health*, *4*, 95–99. https://doi.org/10.1016/j.ssmph.2017.11.008

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, *4*(11), 1208–1214. https://doi.org/10.1038/s41562-020-0912-z

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science*, *11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Upchurch, D. M., Lillard, L. A., Aneshensel, C. S., & Li, N. F. (2002). Inconsistencies in reporting the occurrence and timing of first intercourse among adolescents. *The Journal of Sex Research*, *39*(3), 197–206. https://doi.org/10.1080/00224490209552142

van Buuren, S. (2018). *Flexible imputation of missing data* (Second edition). CRC Press, Taylor & Francis Group.

VanderWeele, T. J., & Hernan, M. A. (2012). Results on Differential and Dependent Measurement Error of the Exposure and the Outcome Using Signed Directed Acyclic Graphs. *American Journal of Epidemiology*, *175*(12), 1303–1310. https://doi.org/10.1093/aje/kwr458

Vannieuwenhuyze, J. T. A., Loosveldt, G., & Molenberghs, G. (2014). Evaluating Mode Effects in Mixed-Mode Survey Data Using Covariate Adjustment Models. *Journal of Official Statistics*, *30*(1), 1–21. https://doi.org/10.2478/jos-2014-0001

Villar, A., & Fitzgerald, R. (2017). Using Mixed Modes in Survey Research: Evidence from Six Experiments in the ESS. In *Values and Identities in Europe* (p. 38). Routledge.

Wadsworth, J., Field, J., Johnson, A. M., Bradshaw, S., & Wellings, K. (1993). Methodology of the National Survey of Sexual Attitudes and Lifestyles.

Journal of the Royal Statistical Society. Series A, (Statistics in Society), *156*(3), 407–421.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis* (2nd ed.). Springer. https://ggplot2-book.org/

Wiggins, R. D., Brown, M., & Ploubidis, G. B. (2017). *A measurement evaluation of a six item measure of quality of life (CASP6) across different modes of data collection in the 1958 National Child Development Survey (NCDS) Age 55 years.* (Working Paper 2017/2; p. 25). Centre for Longitudinal Studies. https://cls.ucl.ac.uk/wp-content/uploads/2017/07/CLS-WP-20172.pdf

Williams, T. C., Bach, C. C., Matthiesen, N. B., Henriksen, T. B., & Gagliardi, L. (2018). Directed acyclic graphs: A tool for causal studies in paediatrics. *Pediatric Research*, *84*(4), Article 4. https://doi.org/10.1038/s41390-018-0071-3

Wright, L. (2023). *Many Models in R: A Tutorial* [Preprint]. SocArXiv. https://doi.org/10.31235/osf.io/azvs4

Yap, S. C. Y., Wortman, J., Anusic, I., Baker, S. G., Scherer, L. D., Donnellan, M. B., & Lucas, R. E. (2017). The Effect of Mood on Judgments of Subjective Well-Being: Nine Tests of the Judgment Model. *Journal of Personality and Social Psychology*, *113*(6), 939–961. https://doi.org/10.1037/pspp0000115