



Understanding Society
Working Paper Series

No. 2014 – 01

March 2014

Measurement effects
between CAPI and Web questionnaires
in the UK Household Longitudinal Study

Jorre T. A. Vannieuwenhuyze & Peter Lynn

Institute for Social & Economic Research, University of Essex, Colchester, UK

Measurement effects between CAPI and Web questionnaires in the UK Household Longitudinal Study

Jorre T. A. Vannieuwenhuyze & Peter Lynn

Institute for Social & Economic Research, University of Essex, Colchester, UK

Non-technical summary:

Because of rising costs, the idea has been raised to switch the single-mode data-collection design of the UK Household Longitudinal Study (UKHLS) using only Computer-Assisted Personal Interviewing (CAPI) to a mixed-mode data-collection design using CAPI and Web Self-Interviewing (WSI). The reasoning behind this mixed-mode design is that data-collection by Web is much cheaper than personal interviews because it avoids expensive interviewer wages. As a result, total survey costs should reduce.

However, notwithstanding a possible advantage of lower costs, a mixed-mode design may damage data-quality because of measurement effects between both modes. Measurement effects refer to differences in measurement error accompanying both data-collection modes and may lead to wrong scientific conclusions. The occurrence of measurement effects between modes would not only mean that data of different respondents cannot simply be compared within a UKHLS wave, but also that data across different waves cannot be compared.

Because of the risk of increased measurement error in a mixed-mode UKHLS design, it is imperative to first research measurement error and measurement effects between WSI and CAPI in greater detail before a final decision is made about moving to a mixed-mode design in the UKHLS. Such research must focus on three topics. First, research should try to identify which UKHLS questions are affected by measurement effects between WSI and CAPI. Second, research should investigate how measurement effects can be reduced or even avoided on sensitive questions by optimizing question formulation, questionnaire lay-out, questionnaire delivery, advance letters or calls, reminders, or incentives. Third, research should investigate whether measurement effects can be estimated for affected items where design optimization does not offer solutions. All three research topics may be challenging to research, and it should be understood that there is a risk that for some items/estimates it will not prove possible to theoretically discover mode effects, to avoid mode effects through survey design, or to estimate mode effects through analysis.

Measurement effects between CAPI and Web questionnaires in the UK Household Longitudinal Study

Jorre T. A. Vannieuwenhuyze & Peter Lynn

Institute for Social & Economic Research, University of Essex, Colchester, UK

Abstract: In order to lower costs, the idea has been raised to use a mixed-mode design for the UK Household Longitudinal Study (UKHLS) using Computer-Assisted Personal Interviewing (CAPI) and Web Self-Interviewing (WSI). However, a mixed-mode CAPI-WSI design may damage data-quality because of measurement effects between both modes. Before a decision is made about moving to a mixed-mode design in the UKHLS, measurement effects should be researched in greater detail by identifying questions sensitive to measurement effects, investigating survey design optimization for avoiding measurement effects, and investigating estimation techniques for measurement effects.

Keywords: Mixed-mode survey, CAPI, Web Self-Interviewing, Mode effects, Measurement effects, Unimode design, Instrumental variable, Covariate adjustment, UKHLS

JEL classification: C42, C81, C83

Contact: jtavan@essex.ac.uk (Jorre Vannieuwenhuyze) or
plynn@essex.ac.uk (Peter Lynn).

Acknowledgements: This study was funded by Economic and Social Research Council Award no. ES/H029745/1, "Understanding Society and the UK Longitudinal Studies Centre" (Principal Investigator: Nick Buck).

Measurement effects between CAPI and Web questionnaires in the UK Household Longitudinal Study

Jorre T. A. Vannieuwenhuyze & Peter Lynn

Institute for Social & Economic Research, University of Essex, Colchester, UK

Because of rising costs, the idea has been raised to switch the single-mode data-collection design of the UK Household Longitudinal Study (UKHLS) using Computer-Assisted Personal Interviewing (CAPI) to a sequential mixed-mode data-collection design using CAPI and Web Self-Interviewing (WSI) via self-administration questionnaires (also known as SAQ's). In such a mixed-mode design, sample members are first asked to complete the survey by WSI, and those who do not respond within a reasonable time are afterwards approached by an interviewer at home for CAPI in order to reduce nonresponse. The logic behind this sequential WSI-CAPI mixed-mode design is that data-collection by WSI is much cheaper than data-collection by CAPI because it avoids expensive interviewer wages. As a result, provided that the development costs of WSI are sufficiently small and that a sufficient amount of respondents respond by WSI, total survey costs should reduce while the target sample size remains fixed.

However, notwithstanding a possible advantage of lower costs, a mixed-mode WSI-CAPI design may damage data-quality because of measurement effects between both modes (de Leeuw, 2005; Dillman et al., 2009b). Measurement effects refer to differences in meas-

urement error accompanying both data-collection modes. They thus occur if respondents would respond differently when they completed the survey by CAPI or WSI. As a result, they may seriously affect data analysis outcomes and may lead to wrong scientific conclusions. Indeed, the occurrence of measurement effects between WSI and CAPI within the UKHLS would not only mean that data of different respondents cannot simply be compared within a wave of data-collection, but also that data across different waves cannot be compared (Couper, 2011). Measurement effects might thus be a serious flaw in data quality of the UKHLS.

Because of the risk of increased measurement error in a mixed-mode UKHLS design, it is imperative to first research measurement error and measurement effects between WSI and CAPI in greater detail before a final decision is made about moving to a mixed-mode design in the UKHLS. Nonetheless, a systematic overview of measurement error and measurement effect within a mixed-mode survey in the UKHLS is lacking. This gap is filled by this paper.

Research about measurement effects in an UKHLS mixed-mode design must focus on three topics. First, research should try to identify which UKHLS questions are affected by measurement effects between WSI and CAPI.

Second, research should investigate how measurement effects can be reduced or even avoided on sensitive questions by optimizing question formulation, questionnaire lay-out, questionnaire delivery, advance letters or calls, reminders, or incentives. Third, research should investigate whether measurement effects can be estimated for affected items where design optimization does not offer solutions. All three research are successively discussed in the first three sections of this paper. Finally, Section 4 concludes the paper with a short critical discussion all three research topics because there is a risk that for some items/estimates it will not prove possible to theoretically discover mode effects, to avoid mode effects through survey design, or to estimate mode effects through analysis.

1 Theory

Measurement effects refer to differences in measurement error caused by the data-collection modes, and measurement error may occur during three different phases of a survey measurement process (see Table 1). These phases include the design phase, the response phase, and the administration and analysis phase. In all three phases, measurement effects between WSI and CAPI data-collection may be explained by different characteristics of both data-collection modes (Couper, 2011; de Leeuw, 2005, 2008). First, both modes differ from each other regarding the degree of interviewer involvement. In contrast to CAPI, WSI does not require any interviewer to be involved so that no interviewer effects on responses are expected. Second, both modes differ regarding media-related

factors like the pace and flow of the data-collection process. With CAPI, the pace and flow of the answering process are mostly controlled by the interviewer while they are completely under the respondent's control with WSI. Third, both modes differ regarding factors influencing the information transmission like the presentation of the information. With WSI, questions are usually presented merely visually while with CAPI they can be presented both visually and aurally. All these differences between WSI and CAPI may cause differential measurement of responses to the UKHLS question items. Nonetheless, it should be noted that it remains unclear how these differences translate to exact differences in measurement error between both modes because research on WSI and CAPI comparisons is still scarce (Couper, 2011; Martin, 2011).

1.1 *Measurement effects occurring in the design phase*

The first phase of a survey measurement process involves the design of the survey. This phase depends on decisions made by the research team, and it includes three steps where measurement effects between WSI and CAPI may be introduced. These steps are the formulation of the question, the questionnaire construction, and the transmission of the questionnaire to the respondent.

The formulation of the questions is the first step of a measurement process in which measurement error can be introduced. Measurement error can be introduced here if the formulated questions do not match with the operational definition of the variable of interest. As a consequence, measure-

Table 1. Measurement effects between modes may occur during three different phases of a survey measurement process.

Error	Description
Design phase:	
1) Formulation of the questions	The question formulation may elicit responses that do not match the operational definition of the variable of interest.
2) Questionnaire construction	The influence of, among others, question order, questionnaire lay-out, additional explanation, or the questionnaire introduction on the outcome.
3) Questionnaire transmission	Effect of the communication medium on the outcome (e.g. interviewer effect).
Response phase:	
4) Question comprehension	A mismatch between the respondents' interpretation and the researchers' intended interpretation of the questions and all answer categories.
5) Retrieval of relevant information	Respondent puts insufficient effort into getting the correct answer (i.e. satisficing).
6) Judgement of answer	The respondents maps his answer on the wrong answer category on the response scale
7) Response reporting	Respondent decides to misreport or inaccurately responds the answer.
Administration and analysis phase:	
8) Response administration	Inaccurate administration of the responses by the research team.
9) Timeliness of measurement	A respondent changes between the start of the survey and the tabulation of the results

ment effects between modes can be introduced when questions are formulated differently across the modes (Dillman et al., 2009b). As will be discussed later, it is fairly easy to formulate most survey questions equivalently for CAPI and WSI because both modes can rely on a visual channel of communication for presenting the answer categories (using show cards with CAPI). Nevertheless, some problems may occur if questions are reformulated for WSI in order to facilitate respondents' understanding of the questions.

Once all questions are formulated, the researcher needs to combine the questions into a questionnaire. The

design of this questionnaire depends on important decisions about the question order, the lay-out, the presence of additional explanation, or the questionnaire introduction. The choice for a particular design may affect measurement error because, for example, preceding and subsequent questions can affect respondents' inferences about intended meaning of a particular question and rating scale (Schwarz et al., 2008). Different modes generally require different questionnaire design and may thus lead to measurement effects. With WSI, for example, respondents can go back and forth between questions (at least between questions listed on the same

Web page) and may adapt answers because the locus of control is completely situated on the respondent's side (de Leeuw, 2008). With CAPI, in contrast, the locus of control is situated on the interviewer's side, which makes it impossible to the respondent to adapt answers to subsequent questions.

Once the questionnaire is designed, it needs to be transmitted from the research team to the respondent. This transmission process requires an interviewer when CAPI is used but not when WSI is used. The interviewers, however, may cause measurement error because they may steer the meaning of a question in a particular direction due to their own (mis-)interpretation of the question and response categories, or because of particular interviewer attitudes and behaviours (Loosveldt, 2008). Also, the interviewers' intonation and emphasis on words can lead to different interpretations of questions (Revilla, 2010). Such differences in interpretations caused by interviewers are absent with WSI where the respondent reads the question him- or herself. As a result, measurement effects may occur between WSI and CAPI.

1.2 Measurement effects occurring in the response phase

The second phase of a survey measurement process involves the process of respondents answering the questions. This phase depends on the respondents, and it includes four important steps where measurement effects between WSI and CAPI may be introduced. These steps are comprehending the questions, doing cognitive effort for retrieving relevant information to answer the question, judging the answers, and reporting the answer.

Once the respondent received the questionnaire, he or she must understand the meaning of the questions. Here, measurement error may occur if there is a mismatch between the interpretation of the respondent and the intended interpretation of the researcher with respect to the questions and all possible answer categories. With CAPI, interviewers can use nonverbal cues, can monitor and react to respondents' nonverbal expressions, can answer respondents' questions, solve problems, and give additional information to enhance respondents' comprehension (de Leeuw, 2008; Martin, 2011). Nonetheless, interviewers are also in control of the pace of the interview and may not give respondents sufficient time to consider and understand questions thoroughly (de Leeuw, 2008). Also, clarifying questions cannot be done in a directive manner and may thus be prone to interviewer effects (Loosveldt, 2008). With WSI, in contrast, respondents control the pace themselves but lack interactive help of an interviewer (Klausch et al., 2013).

Once the question has been understood, respondents have to dig in their mind or consult external sources to find correct information for answering the questions. Here, measurement error can be introduced when respondents put insufficient effort into getting sufficient and correct information, and such behaviour is called satisficing (Krosnick, 1991). Satisficing may result in an indiscriminate use of 'yes' or 'agree' responses (acquiescence), the indiscriminate use of one point on a response scale for a range of different items (non-differentiation), or choosing 'don't know' response categories (Jäckle et al., 2010). On the one hand, WSI respondents may be less inclined to expend the

necessary effort compared to CAPI respondents because they might be engaged in different activities while answering the questions (multitasking) and a motivating interviewer is absent (Heerwegh, 2009; Holbrook et al., 2003). Moreover, WSI is more demanding for respondents than CAPI because they require some level of computer literacy apart from the general ability to read, recognize numbers, and know-how to use response formats, while CAPI respondents just listen to the questions and answer orally (Heerwegh and Loosveldt, 2008). On the other hand, CAPI respondents might act in line of a social norm to agree with the interviewer as a guest, especially if the interviewer administers the questionnaire too fast to allow thorough cognitive processing (de Leeuw, 2005).

Once the respondents have gathered the required information for answering the questions, they have to summarize this information in a final judgement by mapping it on the possible answer categories or by fitting it within the answer format (Krosnick, 1991). With WSI, primacy effects may occur if respondents focus on the first response categories and ignore the latter categories (Dillman and Christian, 2005). If no show cards are used with CAPI, recency effects may occur if respondents can not remember the first response categories (Krosnick and Alwin, 1987).

Once the respondent selected the correct answer in his mind, he needs to report this answer. Here, measurement error can be introduced when the respondent decides to misreport his answer because of social desirability (de Leeuw et al., 2008a). Social desirability bias arises if respondents deliberately or unconsciously select the more

socially desirable response in order to portray themselves in a more favourable light than revealing the true answer would achieve (Jäckle et al., 2010). Because the social distance is larger for WSI than CAPI, it is assumed that CAPI engenders more socially desirable responses because WSI is perceived as more private (Heerwegh, 2009). Additionally, measurement error may also occur here when respondents do not put sufficiently effort in responding their complete answer. Especially unstructured or open questions are not suitable for WSI because Internet respondents are more impatient than they are with CAPI (de Leeuw, 2008). As a result, answers to open or unstructured questions may be reported inaccurately with WSI.

1.3 *Measurement effects occurring in the administration and analysis phase*

The third phase of a survey measurement process involves the administration and analysis of the respondents' answers. This phase again depends on decisions made by the research group, and it includes two important steps where measurement effects between WSI and CAPI may be introduced. These steps are the administration of the responses and making conclusions beyond the time frame of the data collection (timeliness). Data analysis can also be a source of measurement error by using incorrect data-analysis models, but these do normally not depend on modes of data-collection.

Once the respondent provided his answers, these answers need to be recorded. With WSI, this recording occurs automatically and is usually not susceptible for measurement errors if no Web design errors were made (de

Leeuw, 2008). CAPI, in turn, relies on the interviewer for recording and the interviewer might affect the responses. For example, the recorded answer can be the interviewer's interpretation of an inadequate answer (Loosveldt, 2008). If respondents give inadequate answers, interviewers must probe, and probing cannot be done in a directive manner and might lead to interviewer effects. This problem might especially occur with field coding, where the interviewer is asked to assign the respondent's reply to an open question to one of several response categories. Additionally, interviewers can easily make typing errors while recording the responses.

Surveys always take some time and respondents' characteristics, attitudes, or status may change during this time (Fricker and Schonlau, 2002). Measurement error can be introduced if there is a correlation between the moment of response and respondent characteristics. WSI respondents usually respond within a short time interval after a contact is made, which allows for faster re-contacts. For CAPI respondents, in contrast, this time interval can be longer because CAPI requires extensive interviewer travel and interviewers can survey only one respondent at a time.

2 Survey design

A first approach to deal with measurement effects tries to prevent measurement effects by optimizing the survey design (Couper, 2011). Such an optimal design must pursue a consistent meaning of questions and responses across different modes (Dillman et al., 2009b; Martin et al., 2007). This pursuit of consistency, however, does not mean that

all respondents should be presented the same questions and the same questionnaire design. In many situations, a rote repetition of questions in both CAPI and WSI is unlikely to achieve consistent responses across both modes because it would avoid measurement effects due to different question formulation and questionnaire design but may introduce measurement effects caused by differences in the other steps of the measurement process. Put differently, design optimization for avoiding measurement effects involves a trade-off between different sources of measurement effects.

An often used example of the pursuit for consistency is the Universal Presentation developed by Martin et al. (2007) and also discussed by Dillman et al. (2009b) and Couper (2008). The Universal Presentation is a set of guidelines for designing questionnaires in a mixed-mode context under the main idea that differences in question design across modes should be avoided as much as possible. Otherwise, if differences are unavoidable, they should be defended by grounded theoretical arguments.

Note, however, that the Universal Presentation presumes that application can take place from the initial planning stages of a survey. The UKHLS does not offer this capability because it is a pre-existing survey and it has already been designed and implemented as a CAPI survey with a self-completion component. The modules and questions carried on the survey have thus not been designed in ways that are always compatible with WSI. The task thus is to propose alternative question and questionnaire designs for WSI that maximize consistency of meaning relative to the already existing CAPI ques-

tionnaire (Dillman et al., 2009b; Jäckle et al., 2010).

2.1 Question construction

The first guideline of the Universal Presentation says that every effort should be made to maintain the same question construction when transferring questions from the CAPI questionnaire to the WSI questionnaire. Maintaining the same question construction implies the use of equal question wording, punctuation, and selective emphasis. Variations in wording, punctuation, or selective emphasis may sometimes be used but should always be based upon evidence that it is more likely to produce consistent responses.

Small differences between both the CAPI and WSI questionnaires might sometimes be required in order to facilitate interviewers' reading out loud with CAPI or respondents' silent reading with WSI. For example, it should be investigated whether the methods used for emphasis in CAPI (boldface, underlining, capitalization, italics, etc.) can straightforwardly be copied to WSI where the respondent him- or herself reads the questions instead of the trained interviewer.

Nevertheless, it should be noted that some of the UKHLS question items are impossible to copy to a WSI questionnaire (de Leeuw, 2008). For example, CAPI allows specific measurements like the cognitive ability test (e.g. memory skills) included in the UKHLS wave three. Such variables are almost impossible to measure without the presence of an interviewer when using WSI. For such questions there is even no alternative question construction possible which would provide consistent responses.

2.2 Response categories

The second guideline says that every effort should be made to maintain the same response format when transferring questions from the CAPI questionnaire to the WSI questionnaire. Maintaining the same response format firstly implies that open questions are not transformed to closed questions or vice versa. It secondly implies that identical response categories and response orders are used for closed questions in both modes.

Some of the open questions in the current UKHLS might be difficult to implement in WSI without modifications. For example, questions that require interviewer coding of stated responses constitute a question type for which a different formulation is inevitable. Within the current UKHLS questionnaire, respondents are asked to report for some items a number of hours or a period of time, but the interviewer records the answer in one of a defined set of categories which are invisible to the respondent (e.g. items *aidhrs* and *paygwc/paynwc/pyuwc*). The interviewer is able to probe an initial response if it does not appear to fit a category, or if the appropriate category is ambiguous. Providing a WSI respondent with a set of explicit categories from which to choose would change the nature of the response task. Alternatively, providing an open-ended text box would be the nearest equivalent to the existing face-to-face response task, but without the possibility of an equivalent to interviewer probing and with an additional substantial post-fieldwork coding task.

The bulk of the closed questions in the current UKHLS employ a response scale that allow only

one response option to be chosen (e.g. *family, memploy, housing, area, stendreas, jbendreas, reasend, jbatt, and jbssectpub*). Here WSI and CAPI may use more or less similar response formats because both modes can rely on a visual channel of communication for presenting the answer categories (using show cards within CAPI). Problems may, however, occur with questions for which a show card includes a relatively long list of response options. Such questions may be susceptible to primacy effects in WSI due to the absence of an interviewer, who can motivate the respondent to think carefully, to take their time, and to re-read all the answer options. Branching the questions may be a solution here, but this may also change the nature of the response task (Malhotra et al., 2009).

Problems may also occur for closed questions with 'select all that apply' options including interviewer's probe for an 'other' unlisted answer (e.g. *servuse, disdif, hcond, hcondn, movy, qualnew, trainpurp, and trqual*). Here, a tempting option for WSI is to request the respondent to explicitly indicate 'yes' or 'no' for each of the response options, but research indicates that this will significantly alter the response distribution (Nicolaas et al., 2011; Smyth et al., 2008, 2006; Sudman and Bradburn, 1982). The best alternative is a layout in which all options are simultaneously visible (not in a drop-down menu), with a tick box for each (Couper, 2008). Additionally, the interviewer's probe for an 'other' unlisted answer can be mimicked by an explicit probe with a white text box at the end of the question, and this probe should be redisplayed until a final explicit 'no other option' is received. Nonetheless, it should be kept in mind that this might not

provide completely consistent answers compared to CAPI.

A particular problem for measurement effects is the presentation of the "Don't Know" option (Klausch et al., 2013). In the current CAPI questionnaire, this option is not explicitly given to the respondents but interviewers are allowed to mark a "dont know" or "refused" option implicitly. It is advised to not show the "Don't Know" option in WSI because doing so would greatly increase the fraction of respondents choosing this option. Omitting the "Don't Know" option might, in contrast, provoke false and random answers. An alternative is to use reactive design features, which, for example, include prompts to the user if an answer is left blank but allow the respondent to continue without completing the question. An experiment with such prompts carried out at IP6 (Al Baghal and Lynn, 2014) suggests that judicious use of these features can reduce item non-response rates to levels similar to those obtained in face-to-face interviews, though it may be possible to use the features only for a limited number of questionnaire items. Additionally, it remains to be shown whether such prompts lead to consistent responses and do not provoke false and random answers either.

2.3 Instructions and explanations

The third guideline says that substantive instructions, help, and examples should be consistent across modes. Transforming CAPI instructions, help, and examples into a WSI questionnaire might, however, be a challenge but, nonetheless, important. Where instructions are thought to be useful in CAPI, they should appear up-front on the web

screen (without any need to roll over or click a link, as this greatly reduces the extent to which they are used). However, this will not always be possible or easy to achieve. An example is the item *paynl*, where the following text appears on-screen in the CAPI instrument:

```
Interviewer Instruction
ENTER TO NEAREST POUND
IF NO DEDUCTIONS MADE PLEASE ENTER 0
IF DON'T KNOW/CAN'T REMEMBER PROBE FOR
APPROXIMATE AMOUNT
RESPONDENT TO CHECK PAYSIP IF POSSIBLE
```

It is hard to conceive how such a complex and multi-faceted set of instructions could be implemented in a functionally-equivalent way in a Web questionnaire. It is very likely that the quality of responses will be lower for WSI respondents as a result.

Transforming examples that are used in the CAPI survey into a WSI questionnaire might also be challenging sometimes. Examples should be presented with WSI in the same way and in the same order as with CAPI. This means that explicit examples in CAPI should also be shown on the screen in WSI, rather than being available only behind hyperlinks. This may, however, lead to large pieces of texts and may enforce satisficing behaviour. If a list of examples is to be 'read if necessary' by interviewers, then hyperlinks might promote equivalent exposure to the material in WSI. That is, if interviewers are instructed to read the examples when they feel the respondent is unsure or confused, and WSI respondents click the examples link when they need help, the context may be similar.

2.4 Questionnaire design

The fourth guideline says that the questionnaire design should be as equivalent

as possible within CAPI and WSI. This implies that question order as well as general preambles to the questionnaire or question modules should be constant. WSI may impose some restrictions here because they can be completed using different technologies, and this may influence how questions and response options appear on the screen (Manfreda and Vehovar, 2008). For example, smart phones may not allow having long questions with long preambles as text may fall off the screen.

Effective questionnaire design in WSI may, however, also ask for grouping questions on similar Web pages or even in response matrices (for example, for the benefits and disadvantages of retirement questions *rtpro1* to *rtpro6*, and *rtcon1* to *rtcon4*). Such designs may simplify question understanding because it eliminates redundancy and requires less effort with keyboard and mouse actions (Manfreda and Vehovar, 2008). Nonetheless, placing items together in a grid may increase item inter-correlations, may lead to more nondifferentiation, and changes the nature of the response task (Heerwegh, 2009).

3 Survey analysis

It is naive to think that all measurement effects can be avoided by proper question formulation and questionnaire design. Indeed, measurement effects caused by differences in, for example, social desirability are hard to avoid by proper survey design. Nonetheless, many UKHLS questions are probably prone to social desirable answers, like questions on income, earnings, health (general health, nutrition, physical activity, smoking), national identity, service use, educational aspir-

ations, and discrimination. The second approach to deal with measurement effects may deal with these forms of measurement effects by trying to estimate and correct for measurement effects within the analysis stage of the survey (Couper, 2011). Unfortunately, it is very difficult to straightforwardly estimate measurement effects because they are completely confounded with selection effects. Selection effects occur when respondents of different modes differ on the variables of interest. As such, selection effects are desired because the absence of selection effects would make a mixed-mode design worthless (Biemer, 2001; de Leeuw, 2005; Dillman et al., 2009a).

Within mixed-mode data, selection and measurement effects are confounded because differences between the groups of respondents selected for the different modes can either be caused by differences between these respondents (i.e., a selection effect) or by differences in measurement (i.e., measurement effects). The simultaneous occurrence of selection effects and measurement effects in UKHLS mixed-mode data would thus complicate evaluation of real data quality and threatens the comparability of the mixed-mode data with data of previous single-mode CAPI waves.

The confounding between selection and measurement effects forms a central topic of the causal inference literature (e.g., among others, Morgan and Winship, 2009; Pearl, 2009; Weisberg, 2010). Indeed, causal inference theory can be applied to mixed-mode surveys because measurement effects merely refer to causal effects of survey mode M on target variable Y . Selection effects, in turn, refer to spurious correlations between Y and M because dif-

ferent population members are selected for the different modes. As a consequence, the relation between Y and M contains both the selection and measurement effects (Figure 1a).

In general, the causal inference literature provides three methods to solve the confounding problem between selection and measurement effects by the inclusion of well-chosen covariates into the analysis model. These methods involve the use of instrumental covariates, confounder covariates, and mediator covariates.

3.1 The instrumental variable model

The first method to circumvent counterfactuals is the *instrumental variable method* (Angrist et al., 1996; Bowden and Turkington, 1990; Heckman, 1996, 1997). The instrumental variable method involves the inclusion of a binary variable I into the analysis model which divides the sample into two groups. However, this variable must meet one important requirement which is that all respondents of one group respond by one single mode, say mode m_1 (Vannieuwenhuyze et al., 2010). Put differently, variable I involves a comparison between a mixed-mode dataset and a single-mode dataset and indicates to which dataset a sample member belongs. As a result, this variable determines the mode of data-collection M and partly breaks the confounding between selection and measurement effects (Figure 1b). The use of an instrumental variable also starts from two assumptions. The first assumption is the *measurement equivalence assumption* and requires that measurement error of mode m_1 is equal in both the mixed-mode and the single-mode sample. The second assumption is the *representative-*

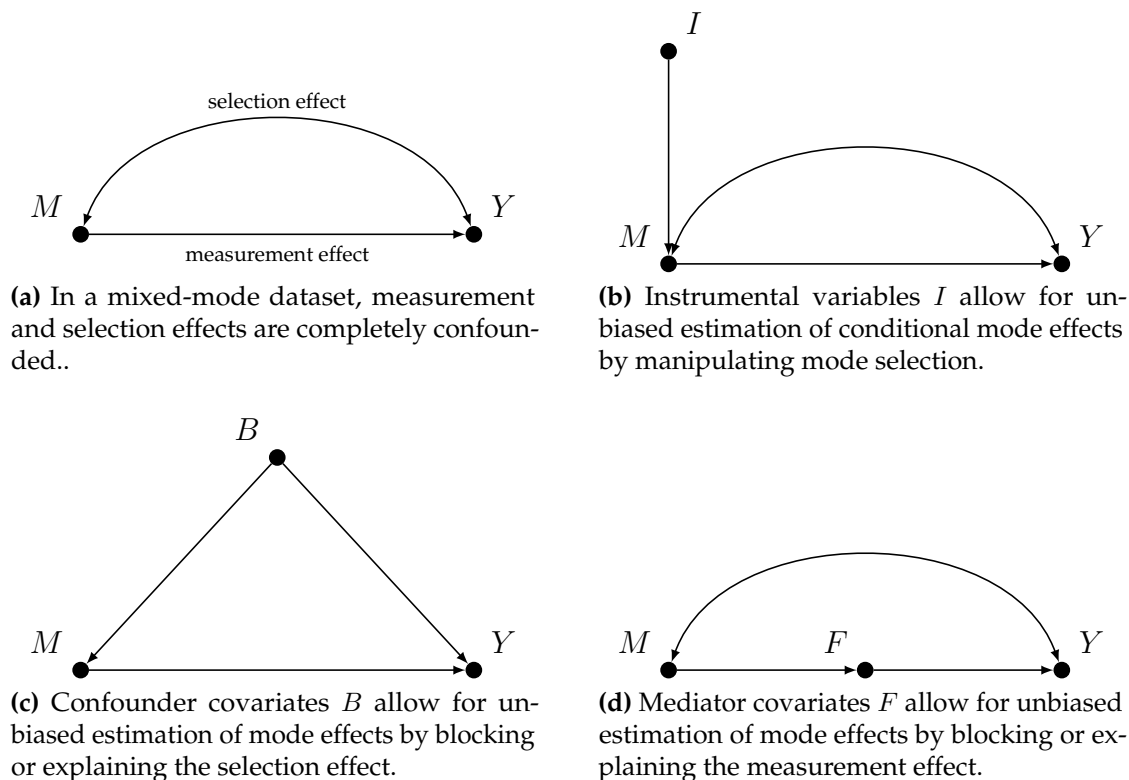


Figure 1. The relations between variables in mixed-mode data can be represented by causal graphs (Pearl, 1995, 2009).

ity assumption and requires that the single-mode and the mixed-mode datasets represent the same population.

The instrumental variable method is not well-known within the mixed-mode literature even though this method merely requires one single-mode dataset and one mixed-mode dataset to be compared. Such two datasets can be obtained within the UKHLS by continuing assigning a subsample of households to the original single-mode CAPI design while the bulk of households is assigned to the mixed-mode WSI-CAPI design.

Nonetheless, even though the instrumental variable method can be used for estimating selection and measurement effects separately, it does not prove itself to be useful for estimation of target statistics, which is usually

the first goal of a survey (Vannieuwenhuyze et al., 2010, 2012). After all, the goal is to use measurement effect estimates obtained by the instrumental variable method for correcting data in the mixed-mode survey which is measured by WSI instead of CAPI. Using the instrumental variable method would mean that such data is imputed on the basis of the single-mode data, or that the mixed-mode data are completely ignored. One may then question why the mixed-mode data are collected at all.

3.2 The confounder model

The second method to circumvent counterfactuals is the *confounder variable method* (Pearl, 1995, 2009; Rubin, 1974, 1978). This method involves the inclusion of a set of variables B into the ana-

lysis model where B explains the selection effects as a common cause of Y and M (see Figure 1c). The confounder variable starts from two assumptions (Morgan and Winship, 2009; Pearl, 2009). The first assumption is the *ignorable mode selection assumption* and requires that B fully captures the selection effect between the modes. If this assumption does not hold, part of the selection effect is not captured and the confounding problem remains. The second assumption is the *mode-insensitivity assumption* and requires that the B variables are mode-insensitive, i.e. there is no measurement effect on B . If this assumption does not hold, part of the measurement effect is channelled through B and the confounding problem remains once again.

Within the existing mixed-mode literature, the confounder variable method has already been widely applied (e.g., among others, Heerwegh and Loosveldt, 2011; Jäckle et al., 2010; Lugtig et al., 2011), but most of these applications merely use socio-demographic variables as confounder covariates. However, such socio-demographic variables might easily be argued to be mode-insensitive, but they might not sufficiently explain why different people are selected for the different modes (Shadish et al., 2008; Vannieuwenhuyze and Loosveldt, 2013; Vannieuwenhuyze et al., 2014). As a consequence, one should not put too much emphasis on socio-demographics for disentangling mode effects. Proper confounder covariates should try to measure population members' capabilities to be contacted and to respond by each mode, population members' mode preferences, and population members' willingness to comprehensively complete surveys (Vannieuwenhuyze et al.,

2014). Questions about mode preferences may thus be good candidates for confounder covariate adjustment. Such questions are already implemented in certain Innovation Panel experiments but need to be investigated further. Other possible confounder covariates may be, for example, respondents' familiarities with computers and Internet use or relative measures of item-nonresponse.

Additionally, data from previous UKHLS waves can also be used as confounder covariates because they can easily be argued to satisfy both the confounder variable method's assumptions. First, variables from previous waves are mode-insensitive because they are all collected by CAPI. Second, it is a pragmatic assumption that selection effects on variables in the mixed-mode wave are well explained by the very same variable measured in the previous waves. Moreover, the different waves also allow for modelling respondents' CAPI response evolution over time which may make the prediction of CAPI responses for WSI respondents even more likely. Nevertheless, assuming that all future UKHLS waves will switch to a mixed-mode WSI-CAPI design, using the initial single-mode CAPI waves might become problematic for modelling CAPI responses of WSI respondents for later waves. Indeed, the larger the gap between a mixed-mode wave and the latest single-mode CAPI wave, the less likely that variables of the single-mode CAPI waves well explain the selection effects between WSI and CAPI. For that reason, it is recommended to look for other covariates which may better allow for disentangling selection effects from measurement effects. Unfortunately, theoretical models and research

towards such covariates are still scarce.

It can also be noted that using the confounder model for correcting measurement effects ignores part of the data. Indeed, correction by confounder covariates within a mixed-mode WSI-CAPI design means that the CAPI responses of the WSI respondents are imputed based on the CAPI responses of the CAPI respondents by matching both groups on the confounder covariate. As a result, such a correction completely ignores the collected responses of the WSI respondents on the target variables. One may then question why data about these target variables data is collected from the WSI respondents, and whether the WSI questionnaire could not have been reduced by only asking for interesting confounder covariates.

3.3 The mediator model

The third method to circumvent counterfactuals is the *mediator variable method* (Pearl, 1995, 2009). This method involves the inclusion of a set of variables F into the analysis model where F explains, in contrast to the confounder variable method, the measurement effect as an intermediate variable between Y and M (see Figure 1d). Like the confounder variable method, the mediator variable method also starts from two assumptions (Morgan and Winship, 2009; Pearl, 2009). The first assumption is the *ignorable measurement assumption* and requires that F fully captures the measurement effects between the modes. If this assumption does not hold, part of the measurement effect is not captured and the confounding problem remains. The second assumption is the *selection-insensitivity assumption* and requires the absence of selection effects on F . If this assump-

tion does not hold, part of the selection effect is channelled through F and the confounding problem remains once again.

Unlike the confounder variable method, almost no application of the mediator variable method can be found within mixed-mode studies so far. The only application of the mediator variable method for the analysis of mode effects in mixed-mode surveys is provided by Vannieuwenhuyze et al. (2014), who used a question about survey liking as a mediator covariate. Other potential mediator variables might be questions about, among others, survey pleasure or survey experience (see for example Loosveldt and Storms, 2008), or variables including information about the number of item nonresponses or primacy and recency effects.

Unlike the instrumental variable method and the confounder variable method, the mediator variable method does use all the data. Indeed, this method imputes CAPI responses of the WSI respondents based on their WSI responses by matching both groups of responses on the mediator covariate.

3.4 Combining the methods

Theoretically, there is no reason to prefer one type of covariates over the other for estimating mode effects in UKHLS data, because the required assumptions are always completely arbitrary. The practical application of all three methods might offer challenges because mixed-mode data fit within the framework of so-called enriched data (Molenberghs et al., 2012). Enriched data require strong and usually empirically unverifiable assumptions (Molenberghs et al., 2012; Verbeke and Molen-

berghs, 2010). It is thus always imperative to carefully assemble the broadest possible evidence towards the assumptions made given the included covariates. Such evidence can be extrapolated from other studies and be included in the analysis of the Understanding Society data as prior information.

Additionally, problematic assumptions of one separate method can be relaxed by integrating another method into the analysis model. As a first example, mediator variables can be used to guarantee the mode-insensitivity assumption of confounder variables. In such an analysis model, the true confounder variables are not observed but a mode-sensitive version is observed instead. The relation between the observed confounder covariates and the mode of data collection can then be blocked by mediator variables. As a consequence, such mediator variables allow for using confounder adjustment with mode-sensitive confounder covariates. As a second example, confounder covariates can be used to guarantee the isolation assumption of mediator covariates. The isolation assumption requires that respondents are more or less randomly allocated to the modes but this assumption is often very implausible. This assumption can be relaxed by matching respondents of both modes on a set of confounder covariates first. As a third example, confounder variables can be used to guarantee the representativity assumption and mediator variables can be used to guarantee the measurement equivalence assumptions within applications of the instrumental variable method. In short, the possibility of complex models provides additional opportunities for estimating selection effects and measurement effects and for finding appropriate confounder

covariates, mediator covariates, and instrumental variables. Indeed, covariates might not perform well when used separately but may do a good job when combined into one analysis model.

4 Discussion

A mixed-mode WSI-CAPI design might be promising for reducing costs within the UKHLS. Nevertheless, such a design comes at the cost of increased measurement error and may make data incomparable over different data-collection waves. For that reason, it is imperative to first research measurement error and measurement effects between WSI and CAPI in greater detail before a final decision is made about moving to a mixed-mode design in the UKHLS. Such research must focus on three topics.

First, there is a need for further theoretical research about measurement error and measurement effects between WSI and CAPI. Such research should try to identify which UKHLS questionnaire items are affected by measurement effects between both modes. Research on this question is already in progress through a report commissioned from NatCen Social Research.

Second, there is a need for further survey-methodological research about optimal survey design in a WSI-CAPI mixed-mode survey. Such research should investigate how measurement effects can be reduced or even avoided on sensitive items by optimizing question formulation, questionnaire lay-out, questionnaire delivery, advance letters or calls, reminders, or incentives.

Third, there is a need for further survey-analytic research about correc-

tion models for mode effects in a WSI-CAPI mixed-mode survey. Such research should investigate whether measurement effects can be estimated for affected items where design optimization does not offer solutions. Such estimation requires the formulation of suitable covariates to be included in the analysis models. However, even if suitable models and covariates could be identified, it is doubtful that the majority of UKHLS data users would be able and willing to use them in their own analyses. As an alternative, data can be corrected and imputed by the Institute for Social & Economic Research before they are released to wider community.

All three research topics may be challenging to research, and it should be understood that there is a risk that for some items/estimates it will not prove possible to theoretically discover mode effects, to avoid mode effects through survey design, or to estimate mode effects through analysis. Nevertheless, the UKHLS program including the Innovation Panel may play a leading role in research on these three topics.

References

- Al Baghal, T. and Lynn, P. (2014). Using motivational statements in web instrument design to reduce item missing rates in a mixed-mode context. under review.
- Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455.
- Biemer, P. (2001). Nonresponse bias and measurement bias in a comparison of face to face and telephone interviewing. *Journal of Official Statistics*, 17(2):295–320.
- Bowden, R. J. and Turkington, D. A. (1990). *Instrumental variables*. Cambridge University press, Cambridge.
- Couper, M. (2011). The future of modes of data collection. *Public Opinion Quarterly*, 75(5):889–908.
- Couper, M. P. (2008). *Designing Effective Web Surveys*. Cambridge University Press, New York.
- de Leeuw, E. (2005). To mix or not to mix data collection modes in surveys. *Journal of Official Statistics*, 21(2):233–255.
- de Leeuw, E., Hox, J., and Dillman, D. (2008a). Mixed-mode surveys: When and why. In de Leeuw et al. (2008b), chapter 16, pages 299–316.
- de Leeuw, E. D. (2008). Choosing the method of data collection. In de Leeuw et al. (2008b), chapter 7, pages 113–135.
- de Leeuw, E. D., Hox, J., and Dillman, D. (2008b). *International Handbook of Survey Methodology*. Erlbaum, New York.
- Dillman, D. and Christian, L. M. (2005). Survey mode as a source of instability in responses across surveys. *Field Methods*, 17(1):30–52.
- Dillman, D., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J., and Messer, B. L. (2009a). Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (ivr) and the internet. *Social Science Research*, 38:1–18.

- Dillman, D., Smyth, J. D., and Christian, L. M. (2009b). *Internet, mail and mixed-mode surveys: the tailored design method*. Wiley, Hoboken, 3rd edition.
- Fricker, R. D. and Schonlau, M. (2002). Advantages and disadvantages of internet research surveys: Evidence from the literature. *Field Methods*, 14(4):347–367.
- Heckman, J. J. (1996). Randomization as an instrumental variable. *The Reviews of Economics and Statistics*, 78(2):336–341.
- Heckman, J. J. (1997). Instrumental variables: A study of implicit behavioral assumptions used in making program evaluations. *The Journal of Human Resources*, 32(3):441–462.
- Heerwegh, D. (2009). Mode differences between face-to-face and web surveys: An experimental investigation of data quality and social desirability effects. *International Journal of Public Opinion Research*, 21(1):111–121.
- Heerwegh, D. and Loosveldt, G. (2008). Face-to-face versus web surveying in a high-internet-coverage population: Differences in response quality. *Public Opinion Quarterly*, 72(5):836–846.
- Heerwegh, D. and Loosveldt, G. (2011). Assessing mode effects in a national crime victimization survey using structural equation models: Social desirability bias and acquiescence. *Journal of Official Statistics*, 27(1):49–63.
- Holbrook, A. L., Green, M. C., and Krosnick, J. A. (2003). Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias. *Public Opinion Quarterly*, 67(1):79–125.
- Jäckle, A., Roberts, C., and Lynn, P. (2010). Assessing the effect of data collection mode on measurement. *International Statistical Review*, 78(1):3–20.
- Klausch, T., Hox, J. J., and Schouten, B. (2013). Measurement effects of survey mode on the equivalence of attitudinal rating scale questions. *Sociological Methods & Research*, 42(3):227–263.
- Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5:213–236.
- Krosnick, J. A. and Alwin, D. F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly*, 51(2):201–219.
- Loosveldt, G. (2008). Face-to-face interviews. In de Leeuw et al. (2008b), chapter 11, pages 201–220.
- Loosveldt, G. and Storms, V. (2008). Measuring public opinions about surveys. *International Journal of Public Opinion Research*, 20(1):74–89.
- Lugtig, P., Lensvelt-Mulders, G. J. L. M., Frerichs, R., and Greven, A. (2011). Estimating nonresponse bias and mode effects in a mixed-mode survey. *International Journal of Market Research*, 53(3):669–686.
- Malhotra, N., Krosnick, J. A., and Thomas, R. K. (2009). Optimal design

- of branching questions to measure bipolar constructs. *Public Opinion Quarterly*, 73(2):304–324.
- Manfreda, K. L. and Vehovar, V. (2008). Internet surveys. In de Leeuw et al. (2008b), chapter 14, pages 264–284.
- Martin, E. A., Childs, J. H., DeMaio, T., Hill, J., Reiser, C., Gerber, E., Styles, K., and Dillman, D. A. (2007). Guidelines for designing questionnaires for administration in different modes. Washington, DC: U.S. Census Bureau.
- Martin, P. (2011). A good mix? mixed mode data collection and cross-national surveys. *ASK Research & Methods*, 20(1):5–26.
- Molenberghs, G., Njeru Njagi, E., Kenward, M. G., and Verbeke, G. (2012). Enriched-data problems and essential non-identifiability. *International Journal of Statistics in Medical Research*, 1:16–44.
- Morgan, S. L. and Winship, C. (2009). *Counterfactuals and causal inference: methods and principles for social research*. Analytical methods for social research. Cambridge university press, New York, (N.Y.).
- Nicolaas, G., Campanelli, P., Hope, S., Jäckle, A., and Lynn, P. (2011). Is it a good idea to optimise question format for mode of data collection? results from a mixed modes experiment. Retrieved 12/11/2013 from <https://www.iser.essex.ac.uk/publications/working-papers/iser/2011-31>.
- Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.
- Pearl, J. (2009). *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York (N.Y.), 2nd edition.
- Revilla, M. (2010). Quality in unimode and mixed-mode designs: A multitrait-multimethod approach. *Survey Research Methods*, 4(3):151–164.
- Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational psychology*, 66(5):688–701.
- Rubin, D. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 6(1):34–58.
- Schwarz, N., Knäuper, B., Oyserman, D., and Stich, C. (2008). The psychology of asking questions. In de Leeuw et al. (2008b), chapter 2, pages 18–34.
- Shadish, W. R., Clark, M. H., and Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? a randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association*, 103(484):1334–1343.
- Smyth, J. D., Christian, L. M., and Dillman, D. A. (2008). Does ‘yes or no’ on the telephone mean the same as ‘check-all-that-apply’ on the web? *Public Opinion Quarterly*, 72(1):103–113.
- Smyth, J. D., Dillman, D. A., Christian, L. M., and Stern, M. J. (2006). Comparing check-all and forced-choice question formats in web surveys. *Public Opinion Quarterly*, 70(1):66–77.

- Sudman, S. and Bradburn, N. M. (1982). *Asking Questions*. Jossey-Bass, San Francisco.
- Vannieuwenhuyze, J. and Loosveldt, G. (2013). Evaluating relative mode-effects in mixed mode surveys: Three methods to disentangle selection and measurement effects. *Sociological Methods and Research*, 42(1):82–104.
- Vannieuwenhuyze, J., Loosveldt, G., and Molenberghs, G. (2010). A method for evaluating mode effects in mixed mode surveys. *Public Opinion Quarterly*, 74(5):1027–1045.
- Vannieuwenhuyze, J., Loosveldt, G., and Molenberghs, G. (2012). A method to evaluate mode effects on the mean and variance of a continuous variable in mixed-mode surveys. *International Statistical Review*, 80(2):306–322.
- Vannieuwenhuyze, J., Loosveldt, G., and Molenberghs, G. (2014). Evaluating mode effects in mixed-mode survey data using covariate adjustment models. *Journal of Official Statistics*, 30(1):1–21.
- Verbeke, G. and Molenberghs, G. (2010). Arbitrariness of models for augmented and coarse data, with emphasis on incomplete data and random effects models. *Statistical Modelling*, 10:391–419.
- Weisberg, H. (2010). *Bias and causation: Models and judgment for valid comparisons*. Wiley, Hoboken, New Jersey.