Informed Design of Mixed-Mode Surveys

Evaluating mode effects on measurement and selection error

Thomas Klausch

Informed Design of Mixed-Mode Surveys

Evaluating mode effects on measurement and selection error

Geïnformeerd Onderzoeksontwerp van Mixed-Mode Surveys:

Evaluatie van mode effecten op meet- en selectiefouten

(met een samenvatting in het Nederlands)

PROEFSCHRIFT

Ter verkrijging van de graad van doctor
aan de Universiteit Utrecht op gezag van
de rector magnificus prof. dr. G.J. van der Zwaan,
ingevolge het besluit van college voor promoties
in het openbaar te verdedigen op vrijdag 10 october 2014
des middags 16:15

door

Lars Thomas Klausch

Geboren op 8 mei 1981
te Halle (Westf.), Duitsland

Promotor: Prof. dr. J. J. Hox

Co-promotor: Dr. J. G. Schouten

Leden van de beoordelingscommissie:

Prof. dr. P. Sturgis
Prof. dr. S. van Buuren
Prof. dr. J. van den Brakel
Prof. dr. P. G. M. van der Heijden
Prof. dr. F. A. van Tubergen

To my parents

Rita, Peter, Malise, and Herbert

List of Contents

List of Tables

List of Figures

1 Introduction

Social researchers have always been concerned about the quality of data collected by one of their most central research methods: sample surveys. In making valid inference about relevant study populations, it is desirable that estimates from surveys are precise. In other words, an accurate survey design minimizes the size of error of final estimates. Unfortunately, the process of conducting a survey – sampling and recruiting individuals, as well as administrating questionnaires – cannot be considered free of error. Therefore, researchers need to know and evaluate which elements of a survey design can impact the size of survey error before implementing a particular survey project at hand.

This PhD thesis is concerned with evaluating the role of the mode of administration – the 'survey mode' – as a central component of any survey design in impacting survey error. The survey mode represents the channel of communication between researcher and respondents, which is applied for recruiting and interviewing individuals. As such, it may potentially have great influence on the accuracy of survey results (De Leeuw, 2008).

Suppose a researcher is interested in the proportion of individuals in a country who consider giving a vote to one of the right-wing extremist parties at the next national elections. A natural way for collecting information on this topic would be asking a random sample of individuals from the population about their voting intentions. In doing so, it has to be decided how to choose a sample of persons, how to contact them, how to pose the question, and how to record a party as an answer.

The survey mode comes into play at each of these stages. For example, in a survey using the telephone as method of communication, interviewers may call randomly

selected telephone numbers and invite respondents to answer some questions on ‘politics’. One of the questions posed may be which party the respondent will most likely vote for at the next elections. However, a likewise legitimate strategy represents sending letters to a random sample of housing addresses, in which an accompanying questionnaire with return envelope asks the same question in writing. Yet another strategy may send out letters with hyperlinks to an online survey asking respondents to access and complete web questionnaires on a personal computer.

In an ideal survey world, it would not matter which of these modes a researcher chooses. For an accurate estimate of extremist party choice, it is important that the sample is a random selection from the population of interest, that all individuals reply, and that the questions are answered correctly. Unfortunately, these quality criterions may be hampered by the mode of administration. For example, as of today, not every person can be reached on landline phones anymore, but many telephone sampling techniques still rely on landline telephone numbers. Also, by far not all sampled individuals can be contacted and are willing to reply. Especially in online surveys, the rate of respondents compared to the number of sampled persons can be very small in practice. Errors may also occur in the answering process. For example, some respondents may not be willing to ‘admit’ a preference for extremist parties to an interviewer, but they might be more willing to do so in the anonymous answering situation of paper or online questionnaires.

In practice, a distinction is made between survey designs in which a survey is administered by a single mode of administration only (‘single-mode surveys’) and designs which apply multiple modes during data collection, so-called ‘mixed-mode surveys’ (De Leeuw, Dillman, & Hox, 2008; De Leeuw, 2005; Dillman, Smyth, & Christian, 2009, pp. 206–310; Groves et al., 2010, pp. 175–177). The move towards mixed-mode data collection designs is motivated by a need in survey research to increase population coverage of survey designs and raise response rates of single-mode designs while saving on costs.

An example of a mixed-mode survey is a design that combines modes of data collection ‘in sequence’. This strategy is based on the idea to first offer a single-mode as a response channel (e.g., online questionnaires). Subsequently, nonrespondents under

this mode are provided with the chance to reply in a second (e.g., telephone) or even a third mode (e.g., in person interviews). In doing so, more respondents can typically be reached. In addition, sequential mixed-mode designs can start with an inexpensive mode (e.g., online) and only later use more expensive modes (e.g., telephone or face-to-face). This procedure may lead to cost efficient designs. However, even though often more respondents can be recruited for the mixed-mode survey, there is a chance to collect less accurate answers under some modes. In the 'voting' example given above, combining a web with a telephone survey might increase overall response rates. However, if respondents on the phone are less willing to admit their intention to vote for an extremist party, survey error may be increased despite the higher response gained by the sequential mixed-mode survey.

Fostered by the growing number of available types of modes in recent years, choosing modes for mixed-mode designs has become a key decision problem in contemporary survey research. In the course of this dissertation, we review, develop, and apply methodology to evaluate empirically, whether particular combinations of survey modes in a mixed-mode survey may be beneficial for survey error and which risks may be involved in taking different design choices. This information can help survey practitioners to take informed decisions about combinations of modes when designing mixed-mode surveys.

To introduce the reader to the relatively large field of mixed-mode research the following two sections give an overview on the relation of survey modes, mixed-mode survey designs, and survey error (sections 1.1 and 1.2). Subsequently, general research objectives and methodological problems in choosing modes for mixed-mode data collection surveys are reviewed (sections 1.3 and 1.4), before the contributions of this dissertation are introduced (section 1.5).

1.1 The Relationship of Survey Mode and Survey Error

In the past decades, survey methodology has seen an upsurge in available modes driven by accelerated technological innovation. Whereas traditional channels included personal interviews ('face-to-face') or paper questionnaires, often mailed to and from respondents ('paper and pencil' or 'mail' modes), 'telephone surveys' became a popular

alternative in the 1970s (Groves, Biemer, & Lyberg, 1988; Groves & Kahn, 1979). In addition, computers supported administration of telephone and, later, also personal interviews. Another important development marked the survey landscape in the second half of the 1990s, when the internet became available as a new communication channel (Berrens, Bohara, Jenkins Smith, Silva, & Weimer, 2003; Buchanan & Smith, 1999; Couper, 2000; De Leeuw & Hox, 2011; Tuten, Urban, & Bosnjak, 2002). Web interviewing quickly revolutionized survey research, facilitating the administration of questionnaires to large groups of individuals at a never seen pace, shortening fieldwork times and increasing timeliness of survey results (Couper & Miller, 2008). The web channel is currently quickly diversifying further, from the original administration on desktop PCs to mobile electronic devices, such as smart phones or tablet PCs (Toepoel & Lugtig, 2014).

Today it is generally accepted that the major types of modes are face-to-face (in person), telephone, paper-based (mail), and online (web). However, the technological development has made available a number of sub-types (Table 1.1). These differ by several factors: interviewer presence (yes or no), question presentation to respondent (oral or visual), response channel for communicating the answer back (oral, on personal computer [PC], telephone, or mobile device), and how communicated answers are entered as data (paper or PC-based questionnaires completed by interviewers or respondents, voice recognition, touchtone entry, or touchscreen). Further details and alternative classifications are provided by Tourangeau, Rips, and Rasinski (2000, pp. 290–293), Biemer and Lyberg (2003, pp. 189–205), and Groves and colleagues (2010, pp. 150–153).

Independent of mode, any survey design is characterized by a series of successive stages that involve the definition of a sampling frame, selecting a sample, approaching individuals (contact and consent), administering the survey (answering), analysis, and reporting (Groves et al., 2010, p. 47). On each of these stages a survey design may accumulate particular kinds of error, which are summarized collectively in the Total Survey Error (TSE) framework (Biemer & Lyberg, 2003, pp. 34–38; Biemer, 2010a, 2010b; Groves & Lyberg, 2010; Groves, 1989; Lyberg, 2013). This framework defines TSE as the deviation of a sample estimator from the population true score and discusses

multiple reasons why both quantities may not be equal in real-world sample surveys (Table 1.2).

Table 1.1: Taxonomy of main types of single-mode surveys and sub-types with their major characteristics

Main type	Sub-Forms	Interviewer present	Question Presentation	Response channel	Data Entry
Face-to-face	PAPI	Yes	Oral	Oral	Paper by interviewer
	CAPI	Yes	Oral	Oral	PC by interviewer
Telephone	Conv. Phone	Yes	Oral	Oral	Paper by interviewer
	CATI	Yes	Oral	Oral	PC by interviewer
	VRE	No	Oral	Oral	Voice recognition
	TDE	No	Oral	Typing	Touchtone data entry
Paper	Paper-SAQ	No	Visual	On Paper	Paper by respondent
Internet	Web-SAQ	No	Visual	On PC	PC by respondent
	Mobile	No	Visual	On Mobile	Touchscreen by respondent
	Web-Video	(Yes)	Oral/Visual	On PC	PC by respondent

Overview partly based on: Biemer & Lyberg, 2003, p. 189; Groves et al., 2010, pp. 150–153; Tourangeau, Rips, & Rasinski, 2000, p. 292
PAPI: Paper and Pencil Interview; CAPI: Computer Assisted Personal Interview; Conv. Phone: Conventional Phone; CATI: Computer-Assisted Telephone Interviewing; VRE: Voice Recognition Entry; TDE: Touchtone Data Entry; Paper-/Web-SAQ: Self-Administered Questionnaire (via paper / online); Mobile: SAQ-web based on mobile devices (smartphones, tablets); Web-Video: SAQ-web with audio/video

We can distinguish errors due to sampling from errors not caused by the sampling process (non-sampling errors) as well as between random and non-random (systematic) error, the latter also called bias (Biemer & Lyberg, 2003, p. 35). Sampling error is caused by the random sampling process itself, because only a sub-set, but not the whole population is observed in a sample (Table 1.2, row 2). Its contribution to survey error is normally only random variance. However, three major components of non-sampling error are distinguished: selection error, measurement error, and analysis error. Their contributions to survey error are often non-random (Table 1.2, rows 1 and 3 to 5).

In understanding the way a survey mode may impact the TSE, it is important to realize the role of the mode as both a factor impacting the selective stages (sampling

frame, sampling, contact, and consent) as well as the question-answer (measurement) process of a survey design (Table 1.2, last column). Both aspects are discussed in detail in the following two sections.

Table 1.2: Components of total survey error (TSE) across the survey process with examples for differential impacts of mode on error

Stages in Survey Process	Primary Error Component	Error terminology in this book	Potential reasons for mode impact on error
Definition of sampling frame	Coverage (or Frame)	Selection	Erroneous address lists, telephone under-coverage, incomplete internet access, unavailability of email lists
Sampling	Sampling	Sampling	Response rate differences, sample sizes
Contact and Consent	Nonresponse	Selection	Salience of contact mode, ability to impact persuasive power and effort, salience of target variables and survey topics, sensitivity of topic, mode preferences
Administration of Questionnaire	Measurement	Measurement	Differences in question wording and presentation, visual vs. audible information transmission, social situation (presence of interviewer)
Analysis/Reporting	Processing Adjustment Reporting	Analysis	Differences in accuracy of data entry, availability and quality of auxiliary information

Partly based on: Biemer & Lyberg, 2003, p. 39; Dillman et al., 2009, pp. 310–322; Groves & Lyberg, 2010; Groves, 1989, p. 10

1.1.1 The survey mode as a selection method

Determinant for the selective properties of modes is the mode of communication applied during survey administration. Respondents need to be able to use a mode besides being generally willing to accept it for answering questions. These aspects relate to the coverage, contact, and consent stages of the survey process and thus 'selection error' (also referred to as 'non-observation error'; Groves, 1989, p. 10).

Any sample survey is based on a random sample from an enumeration of population units, the sampling frame (Table 1.2, first row). As far as this 'list' of population members does not match the target population, coverage error (also called frame error)

by population ‘under-coverage’ may be a result. The definition of a sampling frame is strongly determined by the eventual mode of communication. Telephone surveys are limited to individuals or households with telephone access (‘telephone sampling frame’). However, over the past decades, landline telephone coverage in many Western countries has decreased (Blumberg, Luke, Cynamon, & Frankel, 2008; Thornberry & Massey, 1988; Trewin & Lee, 1988). In addition, mobile phones have substituted a large share of landline telephone connections giving rise to so-called “mobile-only” populations. This development requires new sampling frames that include landline and mobile populations to increase overall population coverage of telephone surveys (Busse & Fuchs, 2011; Gabler & Häder, 2009; Mohorko, De Leeuw, & Hox, 2013a).

In mail and face-to-face administered modes, address-based sampling frames (ABS) are often central. Also ABS frames may be erroneous, however, if commercial providers have only incomplete information on the number or names of individuals in a certain area or particular individuals are included more than once on a list (Dillman, Smyth, et al., 2009, pp. 46–49). When alternative sampling methods, such as ‘random walk’ procedures, are used in face-to-face surveys, over-coverage of the population may be a problem (inclusion of non-eligible units). National statistical institutes, like Statistics Netherlands, often have access to general address registers of populations (cf. chapter 2 in this book). Such ABS frames are of high quality and cover essentially the whole population (with exception for homeless and illegal immigrants).

As far as the ‘internet’ as a survey mode is concerned, online sampling frames are typically not available. Sometimes it is suggested that email lists are proper sampling frames for web surveys, but email lists of general populations, such as The Netherlands, are certainly not available in practice (and even if they were, many individuals have multiple email addresses leading to multiple inclusion into sampling frames). Therefore, general population surveys using web need to rely on alternative sampling frames, often those associated with the more traditional modes: ABS frames or telephone sampling frames. Statistics Netherlands, for example, uses an ABS frame in surveys involving the web mode. In addition to the coverage error encountered in these frames, incomplete household internet access of a population is regarded as a web-specific coverage problem (Berrens et al., 2003; Blyth, 2008; Couper, Kapteyn, Schonlau, & Winter, 2007; Couper, 2000, p. 200; De Leeuw & Hox, 2011; Kwak & Radler, 2002; Mohorko,

De Leeuw, & Hox, 2013b). The household internet access rate in The Netherlands is high and estimated at 95% as of 2013 (Eurostat, 2014). However, there is considerable variance across European countries and internet under-coverage represents a considerable problem in some countries.

The second source of selection error is the survey nonresponse process (Table 1.2, row 3). Both non-contact and refusal to participate after successful contact are causal to incomplete observation of the sample (unit nonresponse), which may lead to 'nonresponse error'. The role of the mode transcends to these stages. For example, the salience of the contact attempt varies across modes. For successful interviewer contact (face-to-face or telephone), persons need to be at home during the time of contact, so that employed and time-short persons are more difficult to reach (Durrant & Steele, 2009; Groves & Couper, 1998; Lynn & Clarke, 2002; Nicoletti & Peracchi, 2005). Conversely, paper questionnaires or web invitations can be present in the household for extended periods of time and can also be seen at times inaccessible to interviewers.

Persuasion to cooperate after contact is likely to work differently across modes, too. Important persuasive cues are only effective, when interviewers are involved (Dijkstra & Smit, 2002; Groves, Cialdini, & Couper, 1992; Groves & Couper, 1996). For example, interviewers can use tailored strategies to convince potential respondents to participate (Snijkers, Hox, & De Leeuw, 1999). On the phone or in writing such effects might be harder to achieve than in person. Furthermore, interviewers might increase salience of survey topics, which can be problematic when topics are sensitive. In such situations, individuals may anticipate unpleasant questions and refuse to participate or break-off participation during an interview (Adua & Sharp, 2010; Groves et al., 2006; Groves, Presser, & Dipko, 2004; Groves, Singer, & Corning, 2000; Peytchev, 2009; Voogt, 2004).

A direct consequence of mode differences in contact and refusal behaviour is represented by response rate differences. Meta analyses of response rates across surveys show that face-to-face interviews often evoke higher response rates than mail or telephone modes (Hox & De Leeuw, 1994). Furthermore, web-based interviewing seems to evoke lower response rates than the other modes (cf. introduction to chapter 2; Manfreda, Bosnjak, Berzelak, Haas, & Vehovar, 2008; Shih & Fan, 2008). These results

can be related to the stronger persuasive role of the interviewer in achieving contact and consent (e.g., mailed contact attempts are less persuasive than in person contact).

In sum, modes may have a great impact on selection error by influencing the selective stages of the survey process (Table 1.2, rows 1 and 3). By impacting the size of response rates, modes may additionally influence (random) sampling error (Table 1.2, row 2), although the exact size of this error in the response set is controllable by the gross sample size.

1.1.2 The survey mode as a measurement method

The second component of nonsampling error in the TSE framework is 'measurement error' (also referred to as 'observation error'; Groves, 1989, p. 10). Measurement error is incurred during the actual administration of the survey by means of, depending on mode, self-administered questionnaires or interviewers (Table 1.2, row 4).

In principle, measurement error occurs due to imperfections in the question-answer process, which is defined by Tourangeau, Rips, and Rasinski (2000, p. 8) as comprehension of a question, retrieval of relevant information from memory, forming a judgment on the possible answering options, and finally deciding on an appropriate and desired answer. A great number of causes of measurement error are discussed in the literature. The social situation, question complexity, channels of communication, and the motivation and ability of respondents represent important factors in determining the size of measurement error (Krosnick, 1991, 1999). Another source of measurement error is 'specification error' which is caused by invalid operationalization of concepts of interest into survey questions ('construct validity' in psychological literature; Groves & Lyberg, 2010). Part of the literature discusses specification error as a nonsampling error component that is unrelated to measurement error (Biemer & Lyberg, 2003, p. 39; Hox, De Leeuw, & Dillman, 2008).

Since the mode determines the communication environment during administration of a survey, it may greatly impact the situation under which measurement error can occur. The role of the mode in the question-answer process therefore has received great attention in the methodological literature. The following summary distinguishes three central factors that have been identified as potential impacts (cf. also theory section in

chapter 3). More detailed discussions are provided by De Leeuw (2008), Dillman and colleagues (Dillman & Christian, 2005; Dillman & Messer, 2010; Dillman, Smyth, et al., 2009, pp. 310–320; Dillman, 2009) and Tourangeau and colleagues (2000, pp. 289–312).

First, the survey mode sets the social situation under which answering takes place. In particular, it determines whether questionnaires are administered in the presence of interviewers (e.g., PAPI, CAPI, conventional telephone, or CATI) or in a self-administered environment (e.g., paper self-administered questionnaire, web questionnaires). A prominent effect of the presence of interviewers is that of 'social desirability' bias, especially considering questions that are perceived as sensitive or threatening by respondents. This effect is one of the most well documented sources of measurement error bias in interviewer administered surveys, which has been shown to be smaller in the anonymous situation of self-administered surveys that study the same socially sensitive phenomena (Aquilino, 1994; De Leeuw & Hox, 2008; De Leeuw, 2008; Heerwegh, 2009; Kreuter, Presser, & Tourangeau, 2008; Paulhus, 1984; Presser & Stinson, 1998; Sakshaug, Yan, & Tourangeau, 2010). Telephone modes using computer generated voices or recorded interviewers (TDE / VDE modes, cf. Table 1.1) attempt to reduce the social desirability effect by creating an anonymous interviewing situation on the phone (Tourangeau, Couper, & Steiger, 2003).

However, the role of the interviewer in the question-answer process is important in causing further impact on measurement error. Considering that the answering process requires relatively strong cognitive effort by respondents, for example, for retrieving information and judging possible answers and answering categories, the pace of the interview may impact the accuracy of information processing and rapidity of answering, especially when complicated questions are concerned (De Leeuw, 1992, 2008; Dillman, Smyth, et al., 2009, pp. 311–312). Interviewers often increase the pace of surveys leaving respondents less time to think (i.e., the 'locus of control' is on the side of interviewer). Especially on telephone, pauses for thinking are perceived as unpleasant by many respondents and can lead to less thorough consideration of information in memory and answering options. In self-administered surveys, self-chosen pace may improve the question-answer process (i.e., the locus of control is on the side of the respondent). Contrary to these limitations of interviewer surveys, a classical advantage

of the interviewer is its role in administering complicated questionnaires (e.g., many filter questions or questions which may require elaborate explanations to some respondents).

Second, modes differ by the degree of visual and aural information exchange in the question-answer process (cf. Table 1.1). In interviewer administered surveys (PAPI or CAPI) information exchange is primarily audible, whereas, in self-administered surveys, information exchange is primarily visual (De Leeuw & Hox, 2008; De Leeuw, 1992; Dillman, Smyth, et al., 2009, pp. 314–316). In aural communication the cognitive capacity in memorizing, both, the question and answering options while thinking about an answer is relatively high compared to self-administered surveys, which may impact measurement error in interviewer modes, especially when questions are complex (e.g., due to length, type of judgement required, or number of answering options).

Finally, differences in the tradition of questionnaire design across modes are sometimes held responsible for mode differences in measurement error (Dillman & Christian, 2005; Dillman & Messer, 2010; Dillman, Smyth, et al., 2009, pp. 321–322; Dillman, 2009). A prominent example represents the labelling of answering scales commonly used for attitudinal questions. Whereas questionnaire design of mail surveys often provide explicit labelling for all categories, telephone surveys often disregard reading out all answering options with labels explicitly. Such differences in the measurement instrument may result in differences in measurement error bias or variance of attitudinal items, for example. If questionnaires are designed as similar as possible across modes, for example, by using equivalent wording of questions and answering scales (so-called 'unified mode designs'), the dominant factors in causing mode differences in measurement error remain the social situation and visual/aural information exchange (Dillman, Smyth, et al., 2009, p. 326).

1.1.3 Survey mode and analysis error

A sometimes overlooked component of total survey error is error caused by incorrect processing, adjustment, and reporting of statistics (Table 1.2, last row). These types of errors are associated with the analysis and reporting stage of a survey process and can be considered relatively independent of the actual fieldwork and questionnaire

administration (for details on data processing and other analysis errors see, for example, Biemer & Lyberg, 2003, pp. 215–257; Groves et al., 2010, pp. 329–365).

Still, it is possible to attribute modes a potential impact at the analysis stage of a survey project. In nonresponse adjustment of statistics, for example, the availability or quality of auxiliary data may vary across modes. For example, data from web surveys is surveyed in digital form, but paper-based survey methods have to be entered into administrative systems at a later point. Processing errors may thus be higher in paper-based surveys than in web surveys. In the following, the analysis error component is disregarded, for simplicity, and it is focussed on those errors encountered in the fieldwork process, selection and measurement error.

1.2 Types of Mixed-Mode Designs

Traditionally, a single-mode survey is thought of as a survey that uses only a single mode of communication in contacting and interviewing respondents. For example, a telephone survey is limited to individuals or households with telephone access ('telephone sampling frame'). Recruitment involves sampling from telephone numbers, repeated dial attempts, and interviews on the phone. Likewise, in a survey using paper questionnaires, respondents need to be sampled from address lists, invited in writing, and reply by returning questionnaires by mail. In-person interviews can involve 'random walk' procedures for sampling, recruitment in person, followed by face-to-face interviews. And, finally, web surveys can be based on email or other online invitations.

Clearly, it is possible to use more than one mode either in contacting and/or in interviewing respondents. Such designs are generally called 'mixed-mode surveys'. The primary purpose of mixed-mode surveys is to improve the survey response process and reduce particular sources of survey error encountered in single-mode surveys. The term of a 'mixed-mode' design by itself is, however, not clearly defined and may lead to significant confusion about the type of fieldwork and data collection protocol that is described (Table 1.3). For this reason, it is important to structure mixed-mode designs by their major characteristics and purposes (De Leeuw, Dillman, et al., 2008; De Leeuw, 2005; Dillman, Smyth, et al., 2009, pp. 306–310; Groves et al., 2010, pp. 175–

177). Also in the context of this dissertation, only a particular group of mixed-mode designs is addressed.

Table 1.3: Types of mixed-mode designs and their major characteristics

Type of mixed-mode design	No. of Contact Modes	No. of Data Collection Modes	Time Points	No. of Samples
I. Multiple-contact, single-response mode	Multiple	One	One	One
II. Within-interview switch	Single or multiple	Single-Mode, mode-switch for particular questions, same respondents	One	One
III. Mixed-mode data collection	Single or multiple	Alternative modes for different respondents	One	One
IV. Mixed-mode panel	Single or multiple	Alternative single-modes (or mixed-mode data collections III) for same respondents across time	Two or more	One
V. Cross-sectional design-switch	Single or multiple	Alternative single-modes (or mixed-mode data collections III) for different respondents across time	Two or more	Two or more
VI. Comparative mixed-mode	Single or multiple	Alternative modes (or mixed-mode data collections III) for different respondents (often different populations)	One	Two or more

Partly based on: De Leeuw et al., 2008; De Leeuw, 2005; Dillman et al., 2009, p. 307; Groves et al., 2010, p. 176

A key distinction has to be made between mode of contact and mode of data collection. Whereas a survey can involve multiple contact modes (e.g., pre-notification in writing, reminders on the phone) the administration phase may only use a single mode to collect data (e.g., telephone). When multiple modes are combined in the administration phase, mixed-mode data collection designs are created. As explained below these designs are often called either ‘sequential’ or ‘concurrent’. Another aspect of a research design is the number of time points on which data are collected in the survey project. In longitudinal surveys, the same sample of respondents is observed multiple times, whereas in repeated cross-sectional surveys different samples are surveyed across time. In the course of a longitudinal or cross-sectional survey, switches from one mode of administration to a second may occur, for example. The survey can then be regarded as a mixed-mode data collection design across time. In the following sections, the major

forms of mixed-mode designs will be described along these dimensions. The present dissertation, subsequently, focusses on designs that combine modes of administration in data collection, either in sequential and concurrent mixed-mode surveys or for creating time-series data, as explained in sections 1.3 to 1.5.

1.2.1 Multiple-contact single-response mode designs (Type I)

Type I mixed-mode designs are called 'Multiple-contact single-response mode' designs. Such designs make use of multiple contact modes for a survey, which ultimately, however, uses only a single mode in administration. Dillman et al. (2009, pp. 244–246), for example, suggest to use mailed pre-notifications to introduce the purpose of surveys before interviewers' call attempts or in-person visits, which can increase survey response rates (Dillman, 1991). A similar strategy is important for web surveys. Since sampling frames are not available 'online' (e.g., complete lists of email addresses are only available for very special populations such as students or employees), ABS frames are often applied instead by mailing letters that contain an invitation to complete surveys online (Dillman, Smyth, et al., 2009, p. 307). Other examples of mode combinations in the selection process include reminders by telephone where telephone is not used for data collection or interviewers calling respondents to make appointments for in-person interviews (De Leeuw, 2005).

The objective of all of these measures is achieving a reduction in selection error, by reaching more individuals and increasing likelihood of contact and persuasion. It is important to note that many contemporary 'single-mode surveys' actually are often type I mixed-mode designs. A strict single-mode contact strategy is often even regarded as impracticable or infeasible in official statistics (e.g., a face-to-face survey without pre-notification or contact in any other way but in person, or a web survey of a general population). At Statistics Netherlands, for example, all single-mode data collection surveys involve multiple contact modes (e.g., pre-notification by mail in a telephone survey).

1.2.2 Within-interview switch designs (Type II)

Type II mixed-mode designs may be based on a single or multiple mode contact strategy and resemble type I designs in that initially only a single mode is used in administration (e.g., face-to-face). In addition, however, data collection modes are changed within the interview. The primary examples are designs that try to reduce the interviewer impact for socially sensitive questions in face-to-face surveys. In this situation, interviewers ask respondents to complete part of questionnaires in a self-administered mode, usually computer-based (Tourangeau et al., 2000, p. 290). Within-interview switch designs are sometimes referred as an own type of 'mode', which is called computer-assisted self-interviewing (CASI). CASI may additionally involve recorded questions or video elements (audio-/video-CASI). The primary purpose of type II designs is to decrease measurement error bias for particular sub-sets of questions in a survey, such as social desirability bias (De Leeuw, 2005).

1.2.3 Mixed-mode data collection designs (Type III)

Mixed-mode data collection designs represent one of the most important classes of mixed-mode designs (type III), which apply alternative modes to collect data from different respondents in the same sample and the same study period (De Leeuw, 2005; Dillman, Smyth, et al., 2009, pp. 309–310). Presently, two main types of mixed-mode data collection designs are distinguished: 'sequential designs' and 'concurrent designs'. These terms refer to the moment and order in which sample members are offered response through a particular mode.

In sequential designs, all respondents are first offered response through a single mode of administration. The fieldwork design of this single-mode survey step may involve a single- or, more usually, a multiple-mode contact strategy. Following the fieldwork in the single-mode, a second mode is offered to nonrespondents in the first survey step. In addition, the second mode usually is applied for approaching individuals not covered by the first mode, but who are members of the sample drawn for the mixed-mode survey. Often address- or person-based sampling frames, sometimes with matched telephone numbers, are required for sequential designs.

In practice, often more than two modes are combined sequentially. A typical contemporary example of a sequential mixed-mode design represents the American Community Survey (US Census Bureau, 2010), which in its 2010 version used a sequence of mail, CATI, and CAPI interviewing (Figure 1.1, left). Based on an address-based sampling frame (ABS) a first wave approaches households fully in mail. Nonrespondents are followed up in telephone, for those households, where matching of addresses and telephone numbers is possible. A sample of remaining nonrespondents from mail and telephone, as well as non-covered households (households without matched telephone number) is finally approached in person (CAPI). Statistics Netherlands has used a very similar design for the Crime Victimization survey in the time from 2008 until 2011 (Figure 1.1, middle), starting with a web survey where nonrespondents are followed up in telephone (Kraan, Van den Brakel, Buelens, & Huys, 2010). Contrary to the American Community Survey, however, telephone nonrespondents are not followed up in CAPI. The CAPI mode is reserved for individuals not covered by telephone (further examples on sequential designs are provided by De Leeuw et al., 2008; De Leeuw, 2005).

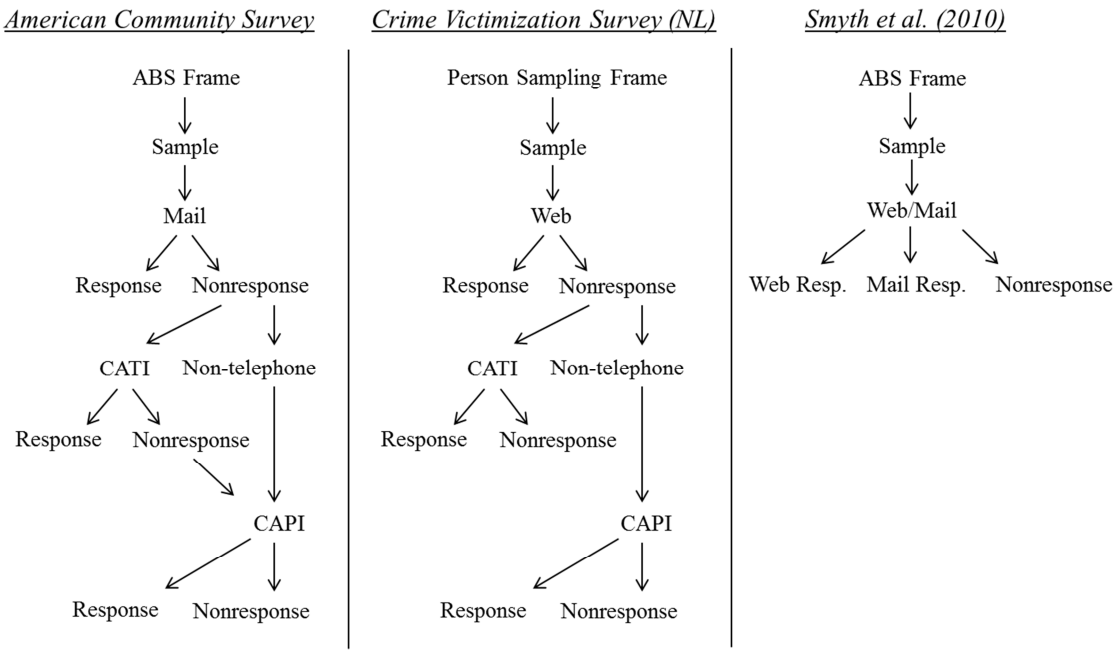


Figure 1.1: Examples of (left) the sequential design of the American Community Survey, the Dutch Crime Victimization Survey (middle), and a concurrent design (right) tested for surveying small towns and communities in the US (Smyth, Dillman, Christian, & O'Neill, 2010)

Contrary to sequential designs, concurrent designs offer multiple modes simultaneously to respondents without explicit ordering. For example, respondents are sent an invitation to a mail survey, but can also call a toll-free telephone number (Dillman, West, & Clark, 1994). Another example represent mail surveys that send paper questionnaires with return envelope, but additionally offer the option to complete a survey online (Figure 1.1, right). By offering multiple modes as response channel it is generally believed that respondents choose a mode which they prefer (Haan, Ongena, & Aarts, 2014; Millar & Dillman, 2011; Smyth et al., 2010).

Sequential and concurrent mixed-mode designs have received increased popularity in the past years fostered by several parallel developments in international survey research. These can be summarized by four major factors: cost pressures, coverage threats, declining response rates, and technological feasibility (Dillman, Smyth, et al., 2009, pp. 300–302).

The cost of survey projects is greatly impacted by the mode used in administration (for a detailed discussion see Groves et al., 2010, pp. 173–174). Face-to-face surveys are typically regarded as very costly due to high variable costs for interviewer salaries and traveling expenses. Telephone modes decrease administration costs by centrally administering surveys from call-centres, but still involve relatively high variable costs due to the employment of interviewers. To the contrary, self-administered modes, such as mail or web surveys fully save on interviewer expenses and are limited to overhead costs (e.g., for questionnaire development, printing or programming, and mailing). Especially web surveys incur only marginal increases in costs when scaled to large samples (paper-based survey data needs to be digitalized causing additional costs, which are avoided in web).

The availability of the inexpensive web mode has been paralleled by strong reductions of survey budgets in official statistics and commercial research institutions. In face of the ‘new’ web mode as an inexpensive option, using more costly modes for general population surveys, such as face-to-face or telephone, requires strong additional arguments. These arguments are found in potentially lower population coverage by web in many countries and lower response rates in web than in other modes. In The Netherlands, the coverage threat is likely to disappear in the near future, as population

coverage approaches completeness (cf. section 1.1.1), but internet surveys often achieve much lower response rates compared to more traditional modes like telephone and face-to-face (Manfreda et al., 2008; Shih & Fan, 2008). In addition, using traditional modes is also challenged by coverage threats (telephone) and poses the problem of higher costs. On a more general level, response rates have been decreasing over the past decades even in the traditionally response-strong modes (De Leeuw & De Heer, 2002).

Sequential mixed-mode surveys can address, both, population under-coverage as well as low response rates (Dillman, Smyth, et al., 2009, pp. 303–305). Whereas single-mode multiple contact designs are limited to multiple modes in the contact phase, mixed-mode data collection designs offer more response channels for administration (cf. Figure 1.1, left/middle). This step can increase survey saliency during recruitment and provides respondents, who are not able or willing to use certain modes, with more response options. Empirical research has demonstrated that sequential mixed-mode surveys can greatly increase survey response rates (for reviews see chapter 2 and De Leeuw, 2005; Dillman, Phelps, et al., 2009; Lynn, 2013). By reducing population under-coverage and increasing response rates, mixed-mode data collection designs may reduce selection error of single-mode surveys, especially in the low-response web mode.

In addition, mixed-mode data collection designs can address the cost problem by sequencing inexpensive before more expensive modes (De Leeuw, 2005; Dillman, Smyth, et al., 2009, p. 302). A mixed-mode survey, for example, can begin with a web survey in order to collect as much inexpensive data from as many respondents as possible. More expensive modes are only used for the nonresponse share of the mixed-mode sample. These steps are followed in the cases illustrated in Figure 1.1: the American Community Survey starts as sequential strategy with a mail survey and the Dutch Crime Victimization survey uses a web survey as first step of the design. In concurrent designs, the same logic is applied to reduce costs, but modes are offered simultaneously. However, response rates in concurrent designs are viewed more sceptically than in sequential design, often not improving response compared to single mode surveys with low response (De Leeuw, 2005) or even performing worse (Millar & Dillman, 2011; Smyth et al., 2010). These findings imply that offering a choice between modes to respondents, so that these may follow their own ‘mode preference’, is not a generally successful strategy for improving survey response.

1.2.4 Mixed-mode panels and cross-sectional mode switches (Type IV and V)

Type I to III designs are the major forms of mixed-mode surveys for single samples and at single points in time. When the study objective requires assessing change in individuals or aggregate changes across time, mixed-mode designs are extended for a longitudinal (type IV) or repeated cross-sectional (type V) perspective.

Type IV designs are termed mixed-mode panels (Dillman, 2009; Lynn, 2013). One sample is observed under different modes of administration across time. Often, for example, a different mode is used for recruiting than for interviewing panel members in later waves of a panel survey. An example of a mixed-mode panel set-up is the Dutch LISS panel, an online panel, which was recruited in a type III mixed-mode concurrent CATI/CAPI procedure (first wave), but all subsequent data collection have been executed online (Scherpenzeel, 2009). Mixed-mode panels can also apply type III data collection in later waves. Lynn (2013) reports on the UK Household Longitudinal Study Innovation Panel, a face-to-face panel survey, in which in a later wave the effect of a switch to a sequential telephone to face-to-face design was evaluated.

Type V designs involve mode switches between survey waves of repeated cross-sectional surveys. Contrary to panel surveys, separate samples are drawn in repeated cross-sectional waves of the same survey project. Usually, a switch to a different single- or mixed-mode design has not been planned at the outset of a cross-sectional survey. Many repeated cross-sectional national surveys have originally been single-mode surveys, but were switched to a different single mode or a mixed-mode strategy at a later point. An example is represented by the Dutch Crime Victimization Survey (CVS). A face-to-face survey upon initiation, the CVS has undergone several redesign stages to different types of sequential mixed-mode designs (Kraan et al., 2010).

Both type IV and type V designs are primarily motivated by cost considerations. As explained in the previous section, decreasing survey budgets have been a central constraint for official statistical institutes, which often administer long-term cross-sectional national surveys like the CVS. Redesign of such surveys to cost efficient mixed-mode data collection designs involving the web, for example, is a priority and needs to consider changes in total survey error caused by the redesign choice. An objective of redesigns is that time series created by the ongoing survey do not suffer

inconsistent breaks by changes in error or error mixtures (i.e., data from the redesign are comparable to data collected under a different design to allow comparisons across time). This aspect gives rise to several research objectives discussed in further detail in section 1.3.

1.2.5 Comparative mixed-mode designs (Type VI)

Type VI mixed-mode designs take a comparative perspective suggesting that different samples are assigned to different single- or mixed-mode strategies (De Leeuw, 2005). Samples are drawn from different populations for comparative purposes, for example across social or ethnic groups, including cross-national comparisons. Application of different modes for different groups may be motivated by differential mode-access in the (sub-) populations under study, for example.

1.3 Research Objectives in Designing Mixed-Mode Data Collection Surveys

In face of the growing importance of mixed-mode data collection designs in contemporary survey research, the present dissertation takes a deeper look into methodological issues arising when taking decisions about the particular survey modes applied during data collection. The following sections introduce three general research objectives that are frequently relevant in practice. All of these objectives are based on the idea that researchers intend to take an informed decision about the type of modes applied in mixed-mode data collection design (type III) or a mixed-mode redesign (type IV or V), such as a mode-switch in a longitudinal survey, before implementing it in practice. Evaluating the objectives may require experimental data collection prior to any implementation. The objectives increase in complexity in practice, however, and evaluating them empirically is associated with various methodological problems (section 1.4). How some of these problems are addressed in the present dissertation is explained in section 1.5.

1.3.1 Objective 1: Evaluate selection error of candidate designs

Besides the general motivation of a reduction in costs, mixed-mode data collection designs (sequential or concurrent) are based on the idea of increasing survey response rates and decreasing under-coverage of the population (section 1.2.3). Implementations of various types of sequential mixed-mode designs have consistently shown that response rates are increased when response weak modes (e.g., web or mail) are followed up by response stronger modes (e.g., face-to-face, telephone). In fact, low response rates and possibly low coverage in web surveys can be considered one important reason for not implementing a single-mode strategy involving web as the only mode of data collection, for example. Such a strategy would be perhaps most inexpensive among all survey designs currently available.

Response (and coverage) rates are in fact well-established quality criterions in survey research (De Leeuw, 1992; Groves & Lyberg, 2010; Hox et al., 2008; Lyberg, 2013). However, their use as indicators for the size of (systematic) selection error may be limited. This fact is supported by the so-called fixed and random response models, in which the size of nonresponse rates can scale the size of nonresponse (or selection) bias upward.

The fixed-response model assumes that contact and participation of an individual in the selection process is predetermined. The model can be extended to include coverage error bias as an element of selection error bias. Ley Y be a survey target variable and S an indicator of selection, i.e. coverage and response ($S = 1$), or non-selection, i.e. non-coverage or coverage but non-contact or refusal ($S = 0$). Consider now the systematic selection error of the estimator of the sample mean $\bar{y}$ (Bethlehem, Cobben, & Schouten, 2011, p. 42; Groves et al., 2006):

$$\begin{aligned} B(\bar{y}) &= E(\bar{y}) - E(Y) = E(Y|S=1) - E(Y) \\ &= Q\big(E(Y|S=1) - E(Y|S=0)\big), \end{aligned} \tag{1.1}$$

where $Q = \sum_i S_i / N$ is the size of the nonresponse stratum (non-coverage and non-response) relative to the size of the population N. Q is equivalent to the nonresponse rate in surveys with complete coverage. Hence, the size of selection error bias depends on the difference in means of responding and non-responding individuals as well as the

selection rate (response rate) Q, diminishing selection error bias (nonresponse bias) in surveys with higher coverage and lower nonresponse. In surveys with complete response and coverage, selection error bias would be zero. However, the exact size of bias is determined by the contrast of means of response and nonresponse groups, which is determinant of the overall size of bias.

The random response model leads to similar conclusions. The model attributes a fixed response propensity to each population member, instead of assuming predetermined response. Let R indicate response in a given sample, then

$$\rho_i := P(S_i = 1|y_i) \tag{1.2}$$

defines the response propensity of population unit i. Expected nonresponse bias against the sampling frame is given by (Bethlehem, 1988, 2002; Groves, 2006):

$$B(\bar{y}) \approx \frac{\sigma_{Y,S}}{\bar{\rho}}, \tag{1.3}$$

where $\sigma_{Y,S}$ is the covariance of response propensities and survey target variable and $\bar{\rho}$ the population mean of response propensities. $\bar{\rho} = N^{-1}\sum_i \rho_i$ can be interpreted as the expected response rate of a given survey design. It can be seen that the response rate may scale $\sigma_{Y,S}$ upwards for higher nonresponse rates, but similarly to the fixed response model, the bias more crucially depends on a second factor, i.e., $\sigma_{Y,S}$. For small association of target variables and response propensities, nonresponse bias will be smaller. Groves (2006) and Groves and Peytcheva (2008) have demonstrated in a meta-analysis that there was only a weak relationship between response rates and nonresponse bias across variables and surveys. Given this observation, response rates may be bad indicators of nonresponse bias.

The random response model may be extended to include non-coverage as a selective stage, if coverage itself is viewed as a random variable. This assumption seems plausible for internet access, because household equipment with internet is a priori a non-predetermined event (independent of the actual sampling frame used). Similarly, telephone access can be regarded a random event, when a different sampling frame is used, such as address-based sampling or random walk. The results of Bethlehem (1988) and Groves (2006) may therefore be transferable to selection error bias and selection

error bias differences between designs. In sequential and concurrent mixed-mode (type III) designs researchers make substantial effort involving additional financial expenses for improving coverage and response rates. Sometimes, it is implicitly assumed in mixed-mode literature that such increase implies reduction in selection error. Given the observations from the fixed and random response model and the empirical results of Groves, this assumption may not be true in practice. It is therefore desirable to evaluate at the design stage of a mixed-mode data collection survey (type III) or a redesign (e.g., mode switch) of type IV or V, whether a mixed-mode design is required for decreasing selection error or whether a mode switch to a different single mode of data collection implies changes in selection error. This entails evaluating additionally, whether single-mode designs (including multiple-contact single-response mode designs, type I) may offer sufficiently low selection error to fully avoid extra effort of type III designs. Objective 1 is addressed in more detail in chapter 2.

1.3.2 Objective 2: Evaluate measurement equivalence between survey modes

The second concern in type III, IV, and V mixed-mode designs are effects of survey mode on the size of measurement error. These effects are called 'measurement effects' or '(pure) mode effects' (Biemer, 1988; De Leeuw, Dillman, et al., 2008; De Leeuw, 1992, 2005; Vannieuwenhuyze & Loosveldt, 2013; Voogt & Saris, 2005). The term of a mode effect is, however, inconsistently used in methodological literature and may also denote effects of mode on selection error (cf. objective 1) or on total survey error (cf. objective 3). The term 'measurement effect' indicates more clearly that an effect on the measurement error of a survey question is denoted.

The reasons why modes may evoke different sizes of (random or systematic) measurement error are discussed on basic level in section 1.1.2 (cf. also chapter 3). We may call the situation when a survey question evokes the same extent of measurement error under two different modes 'measurement equivalence'. To understand the concept of measurement equivalence on basic level we can use a simple measurement model. Let Y^a and Y^b denote outcomes of survey variable Y measured under two modes, A and B, respectively, and let MB^a denote the population level systematic measurement error (bias) of mode A, then:

$$E(Y^a) = MB^a + E(Y). \quad (1.4)$$

Equivalence of measurement error biases across modes suggests that

$$MB^a = MB^b \Leftrightarrow ME \coloneqq E(Y^b) - E(Y^a) = 0, \quad (1.5)$$

where *ME* is the population level mode effect on measurement error bias (in chapter 4, this quantity is introduced as the 'marginal measurement effect'). In principle, similar definitions can be imposed also on equivalence of measurement error variance or on the full distributional level. Measurement equivalence in all cases implies equivalence of measurement errors (or response distributions) and absence of measurement effects.

There are primarily two reasons underlying the requirement for measurement equivalence in mixed-mode research. First, there is general need for comparability of data collected by different modes. In designs that switch modes of administration, for example in type IV and V designs, an effect on systematic measurement error could be falsely interpreted as substantive change across time. More generally, in case of non-equivalence, analyses conducted on mode-specific response data sets can lead to different conclusions. Measurement effects on systematic measurement error can lead to differing estimates of population means and totals (Biemer & Lyberg, 2003), whereas effects on random measurement error can increase the variance of these estimators and attenuate (i.e., bias) relationship estimates (e.g., correlations) to differing extents (Alwin, 2007; Fuller, 1987; Saris & Andrews, 1991). Another problem known from repeated sequential or concurrent mixed-mode surveys is that measurement effects cause instability of time series data, if the size of mode-specific response groups varies across time (Buelens & Brakel, 2014). Second, and related, sometimes measurements from one mode can be considered the 'preferred measurements', also called the golden standard, or 'benchmark measurements' (Biemer & Lyberg, 2003; De Leeuw, 2005). These measurements are considered the 'best' ones available for a survey variable at hand (e.g., smallest measurement error). Under the benchmark assumption, the occurrence of a measurement effect between modes implies that the comparison mode measures a characteristic of interest with larger error. Combining these modes in data collection then suggests that the overall measurement error bias of a mixed-mode design is increased.

In survey practice, it is a central objective in the design of mixed-mode surveys to prevent the occurrence of measurement effects in mixed-mode data collection designs and after mode-switches in time-series data. A large body of literature discusses approaches for reaching this objective by making adjustments to mode-specific questionnaire designs. One approach uses so-called 'unified mode questionnaires' (Dillman & Christian, 2005; Dillman, Smyth, et al., 2009, pp. 326–329). The primary goal of unified designs is keeping the survey stimulus as similar as possible across modes of administration, for example, by using equivalent wording of questions and answering category labels. More recent research suggests that 'cognitive equivalence' is even more important than unifying question appearance (De Leeuw, Dillman, et al., 2008; De Leeuw, 2005). Such questionnaires try to provide the same cognitive stimulus, but would not necessarily use exactly the same wording. Both approaches need to evaluate measurement equivalence of a questionnaire to control the success of a mixed-mode questionnaire design strategy. This endeavour requires, however, that a diagnosis of the problem – estimation of measurement effects – is possible. Approaches for estimating measurement effects (evaluating measurement equivalence) from fieldwork data are discussed in chapters 3 and 4.

1.3.3 Objective 3: Evaluate the trade-off of selection and measurement error

On a more complex level, mixed-mode surveys might offer an 'optimal' trade-off of measurement and selection error in the total survey error. Literature suggests that combinations of modes may compensate for the disadvantages of some modes while exploiting the advantages of another mode (De Leeuw, 2005; Groves et al., 2010, p. 176). The general idea is that the TSE of a mixed-mode design is minimized or brought to an acceptable level given a certain financial budget (Biemer & Lyberg, 2003).

In mixed-mode data collection designs, the final survey estimate is subject to a 'mix' of errors contributed by all of the modes used in the design (cf. Figure 1.1). To see this, consider the a sample mean of a variable observed in a design like the Crime Victimization Survey, the sequential mixed-mode design shown in the middle panel of Figure 1.1:

$$\bar{y}^{mm} = \pi^{web}\bar{y}^{web} + \pi^{tel}\bar{y}^{tel} + \pi^{f2f}\bar{y}^{f2f}, \tag{1.6}$$

were π gives the proportion of respondents in the mode-specific stages of the mixed-mode response sample ($\pi^{web} + \pi^{tel} + \pi^{f2f} = 1$), and $\bar{y}$ the mode-specific sample means, respectively. The TSE of this estimator is a complex composite of mode-specific errors. For example, the weighted systematic survey error of the web response mean in the mixed-mode data set is

$$\begin{aligned} B(\bar{y}^{web}) &= E(Y^{web}|S^{web} = 1) - E(Y) \\ &= \Big(E(Y^{web}|S^{web} = 1) \\ &\qquad - E(Y|S^{web} = 1)\Big) \\ &\qquad + \Big(E(Y|S^{web} = 1) - E(Y)\Big) \\ &= MB^{web} + SB^{web}, \end{aligned} \tag{1.7}$$

where MB^{web} is the measurement error bias conditional on response in the web proportion of the design, and SB^{web} denotes the selection error bias from the true score mean of Y. In the telephone response set we have, however,

$$\begin{aligned} B(\bar{y}^{tel}) &= E(Y^{tel}|S^{web} = 0, S^{tel} = 1) - E(Y) \\ &= \Big(E(Y^{tel}|S^{web} = 0, S^{tel} = 1) \\ &\qquad - E(Y|S^{web} = 0, S^{tel} = 1)\Big) \\ &\qquad + \Big(E(Y|S^{web} = 0, S^{tel} = 1) \\ &\qquad - E(Y)\Big) \\ &= MB^{tel} + SB^{tel}, \end{aligned} \tag{1.8}$$

reflecting that the bias of the telephone follow-up (i.e., nonresponse under web, $S^{web} = 0$, and response under telephone, $S^{tel} = 1$) is a composite of the selection error bias and the conditional measurement error bias in the follow-up telephone mode. Similar decompositions are possible for the CAPI part of the CVS design (Figure 1.1, middle).

Evaluation of the question, whether the full mixed-mode design can actually reduce or might even increase total survey error across its stages are important. However, the number of bias terms grows rapidly with the number of modes in a design and types of survey error that are distinguished. Nevertheless, knowledge on the sources of TSE, for example measurement or selection, is relevant. If a certain mode in the mixed-mode

design contributes large measurement error bias, adaptations of questionnaires might be possible for reducing such impact, for example, but have to be evaluated against positive effects on selection error bias. The trade-off of systematic measurement and selection error in sequential mixed-mode surveys is discussed in more detail in chapter 5.

1.4 Methodological Problems in Evaluating the Objectives

Despite the practical importance of objectives 1 to 3 for taking decisions about mixed-mode surveys, an evaluation in practice often encounters significant difficulties. The most prominent problem is the unavailability of true scores needed in estimating TSE and its components. A second problem is the estimation of measurement effects from mixed-mode data. Measurement effects can be confounded with selection effects in single-mode comparisons and comparisons of mode-specific response sets in sequential or concurrent mixed-mode surveys. A related practical problem is availability of sufficient auxiliary information that could be used for controlling for the confounding problem. These difficulties are introduced before section 1.5 explains the contribution of the present dissertation.

1.4.1 Unavailability of true scores

Despite the fact that survey error usually is defined against true population values, population true scores of survey target variables are, normally, not available in survey practice (Kreuter et al., 2008). Although this insight is relatively trivial – a survey would not be needed, if population true scores were available – evaluation of survey errors is greatly aggravated by this fact (Biemer & Lyberg, 2003, pp. 34–38; Groves & Lyberg, 2010). Put differently, if true scores are available for a survey variable and all individuals in a sample or a population of interest, evaluating research objectives 1 to 3 would not pose substantial methodological difficulties. This goal would essentially only require an implementation of one or several candidate mixed-mode designs. For example, the mixed-mode bias components of the sequential design introduced in section 1.3.3 (objective 3) could easily be evaluated. Using available true scores, all components of systematic survey error could be estimated directly and unbiasedly. Also

a decomposition of variance components of a survey variable into true score variance and random error variance (sometimes called response variance) would be possible. If true scores are available on population level (i.e., not only sample level) the sampling error component may even be specified more precisely.

Based on this idea, case studies of survey error occasionally make available benchmark information from sources external to the survey, such as voting behaviour, GPAs of students, or employment data. These data subsequently form the basis of methodological evaluations of survey error components (Kirchner & Felderer, 2013; Kreuter, Müller, & Trappmann, 2010, 2013; Kreuter et al., 2008; Olson, 2006; Sakshaug et al., 2010; Voogt & Saris, 2005; Voogt, 2004). Such studies apply so-called 'record check data' as substitutes to true scores. A consistent finding of this research is that measurement error bias of sensitive (threatening) questions is higher in interviewer modes, reflecting social desirable responding. The generalizability of these results to other types of questions is weak, however, because measurement and selection error bias appear to be mainly variable-specific phenomena (Biemer, 2001; Sakshaug et al., 2010). Moreover, record-check studies may have substantial methodological problems themselves (Biemer & Lyberg, 2003, p. 291). For example, time periods of record data and surveys may not coincide or administrative records may be erroneous themselves. Also records are often only available for very specific populations, such as students (Kreuter et al., 2008).

1.4.2 Confounding of measurement and selection effects

In evaluating measurement equivalence between modes (research objective 2), mode difference in measurement error (or equivalence) has to be evaluated. This inference step is problematic in practice, because mode-specific response distributions are impacted by both a selection process and a measurement process causing mode-specific measurement and selection error.

1.4.2.1 Single-to-single mode comparison

To understand the problem on a basic level, consider two single-mode surveys in modes A and B. A simple measurement effect estimator might be

$$\widehat{TE} = \bar{y}^b - \bar{y}^a, \tag{1.9}$$

where $\bar{y}^a$ and $\bar{y}^b$ are the response sample means, respectively. This estimator has been called the 'total', 'overall', or 'mode-system' effect (Biemer & Lyberg, 2003; Biemer, 1988). The total mode effect estimator, however, estimates the compound of a mode-difference in measurement error bias and a mode-difference in selection error bias:

$$\begin{aligned} E\left(\widehat{TE}\right) &= \left(E(Y^b|S^b = 1) - E(Y)\right) \\ &\quad - \left(E(Y^a|S^a = 1) - E(Y)\right) \\ &= TB^b - TB^a \\ &= (MB^b - MB^a) + (SB^b - SB^a) \end{aligned} \tag{1.10}$$

where TB^a and TB^b denote the mode-specific total survey error biases of the response mean estimators against the true population mean of Y and MB^a, MB^b, SB^a, and SB^b denote measurement and selection error bias of modes A and B as defined in formula (1.7), thus for mode A (and B likewise)

$$MB^a = E(Y^a|S^a = 1) - E(Y|S^a = 1), \tag{1.11}$$

$$SB^a = E(Y|S^a = 1) - E(Y). \tag{1.12}$$

Using an estimator of a total mode effect for evaluating equivalence of measurement error bias obviously can lead to false conclusions due to its confounding with a mode difference in selection error bias. The differences in measurement and selection error may be defined as measurement (*ME*) and selection effects (*SE*) on the true scores:

$$ME_{TS} \coloneqq MB^b - MB^a \tag{1.13}$$

$$SE_{TS} \coloneqq SB^b - SB^a \tag{1.14}$$

where the index *TS* suggests that, if a true score Y was available for all individuals, the bias terms in (1.10) and mode effects (1.13) and (1.14) could easily be estimated. However, in the practical situation of survey research, when true scores are not available, estimation of the two components becomes practically impossible. To

approach the problem, a first necessary step is represented by substituting the true scores by another observed quantity, such as a 'preferred measurement' from one of the modes. Suppose mode A represents this 'benchmark', then the measurement and selections effects (1.13) and (1.14) simplify after substitution of Y^a for Y as follows:

$$ME_{TS} = E(Y^b|S^b = 1) - E(Y^a|S^b = 1) \coloneqq ME^b \tag{1.15}$$

$$\begin{aligned} SE_{TS} &= E(Y^a|S^b = 1) - E(Y^a|S^a = 1) \\ &\coloneqq SE(Y^a) \end{aligned} \tag{1.16}$$

where ME^b is called the 'conditional measurement effect' for respondents in mode B, and $SE(Y^a)$ denotes the selection effect (difference in selection error bias) on preferred measurements Y^a. ME^b represents the change (increase) in measurement error bias, whereas $SE(Y^a)$ denotes the change in selection error bias by using mode B instead of mode A to measure Y^a (cf. Vannieuwenhuyze & Loosveldt, 2013).

The outcome $E(Y^a|S^b = 1)$ is a so-called 'counterfactual' quantity or 'potential outcome' (Rubin, 1974, 2005). This conjecture suggests that outcomes under mode A for respondents under mode B can never be observed in reality, but it is assumed that they could have been observed, if respondents under mode B would have provided answers under administration of mode A instead. Under certain circumstances, potential outcomes can be estimated, as briefly described below and discussed in more detail in chapters 4 and 5. This step is required to disentangle measurement and selection effects.

1.4.2.2 Relative effects in a mixed-mode data collection design

A similar confounding problem is present in the case of mixed-mode data collection designs. To illustrate, consider a sequential mixed-mode design with two modes, where A is a follow-up to the nonrespondents under B. Likewise in the comparison of two single-mode surveys above, the total mode effect estimator is

$$\widehat{TE}_2 = \bar{y}_2^b - \bar{y}_2^a, \tag{1.17}$$

where $\bar{y}_2^a$ and $\bar{y}_2^b$ are the mode-specific response sample means in the sequential design, respectively, so that $\widehat{TE}_2$ is the total mode effect estimator between the response groups:

$$E\left(\widehat{TE}_2\right) = E(Y^b|S^b = 1) - E(Y^a|S^b = 0, S^a = 1). \quad (1.18)$$

Now, if mode A donates preferred measurements Y^a, as in the example given above, the conditional measurement effect is, again,

$$ME^b = E(Y^b|S^b = 1) - E(Y^a|S^b = 1) \quad (1.19)$$

and the selection effect becomes

$$SE(Y^a)_2 \coloneqq E(Y^a|S^b = 1) - E(Y^a|S^b = 0, S^a = 1). \quad (1.20)$$

Since $\widehat{TE}_2$ denotes a relative comparison of response groups in a mixed-mode design, these conditional mode effects are also called 'relative' effects (Vannieuwenhuyze, Loosveldt, & Molenberghs, 2014; Vannieuwenhuyze & Loosveldt, 2013). Furthermore, ME^b indicates a change of measurement error bias relative to the preferred mode and $SE(Y^a)$ a change in selection error bias in the mixed-mode design (i.e., by the follow-up in mode A). As in the single-to-single mode comparison above, both effects are, however, confounded in the overall relative mode effect and require estimating the potential outcome $E(Y^a|S^b = 1)$ for successful disentanglement of the effects.

1.4.2.3 Estimation of measurement and selection effects

Estimation of conditional measurement (and selection) effects applies approaches available from causal inference theory. This literature has developed multiple methods for evaluating potential outcomes in the context of inference about treatment effects in non-randomized experimental designs (Morgan & Winship, 2007; Pearl, 2009; Rubin, 1974; Schafer & Kang, 2008). An important assumption is so-called unconfoundedness conditional on exogenous auxiliary variables *X*, also called ignorable treatment assignment (Imbens, 2004; Rosenbaum & Rubin, 1983):

$$(Y^a, Y^b) \perp M \mid X. \quad (1.21)$$

This notation suggests that the potential outcomes Y^a and Y^b are conditionally independent from the so-called selection mechanism *M* into mode-specific response groups in the mixed-mode design. *M* is a general notation to denote any relative

selection mechanism of a mixed-mode data collection design into response groups, where above we used explicit notation using response mechanisms S^b and S^a for a sequential design with two modes or a single-to-single mode comparison (M can be thought of as an indicator specifying the group of respondents who reply using a particular mode *m*).

In the causal inference literature, the auxiliary variables *X* are sometimes referred to as 'backdoor' variables or as confounders of potential outcomes under treatment and control with the treatment assignment mechanism (Morgan & Winship, 2007; Pearl, 2009; Rubin, 1974, 2005). Under the unconfoundedness assumption, conditional measurement effects can be estimated via parametric and non-parametric estimation of the potential outcomes not observed in mixed-mode data collection designs, for example by propensity score weighting (Rosenbaum, 1987), matching (Morgan & Harding, 2006; Rosenbaum, 2002, 2010; Stuart, 2010), imputation (Kang & Schafer, 2007; Schafer & Graham, 2002; Schafer & Kang, 2008), or regression estimation (Imbens, 2004).

1.4.2.4 Estimation of marginal measurement effects

The decomposition of conditional measurement and selection effects, discussed above, takes on a relative perspective between response groups in single-to-single mode comparisons or in a sequential design. Another option is to estimate marginal (population level) measurement effects, equation (1.5). These effects do not condition on a response mechanism and thus require extrapolation to the population. In the single-to-single comparison introduced above, $E(Y^a)$ and $E(Y^b)$ can be estimated using nonresponse adjustment of $\bar{y}^a$ and $\bar{y}^b$ by, for example, calibration weighting to the population (Bethlehem, 2002) assuming missing at random data, MAR (Little & Rubin, 2002; Rubin, 1976) given auxiliary information *X* on sampling frame level:

$$(Y^a, Y^b) \perp (S^a, S^b) | X. \tag{1.22}$$

1.4.2.5 Limitations of estimation approaches

Under both estimation perspectives, a crucial role is played by auxiliary variables *X*. In practice, however, an important problem is the availability of such auxiliary information. One explicit assumption is that *X* is exogenous suggesting that it may not underlie measurement effects itself (i.e., it may not be part of the set of all $\{Y^a, Y^b\}$). A safe indication of exogeneity is that variables are available from external sources, such as registers (sampling frame information). Otherwise it has to be plausibly assumed that auxiliary variables are not affected by measurement effects when measured within a mixed-mode design (chapter 4 in this book; cf. Vannieuwenhuyze, Loosveldt, & Molenberghs, 2014; Vannieuwenhuyze & Loosveldt, 2013).

In practical circumstances, auxiliary data is often limited to socio-demographics or regional information about respondents and nonrespondents. Socio-demographics can be considered relatively robust against measurement effects in mixed-mode surveys or are available from registers. However, the available information is often merely weakly related to response mechanisms (M or S) and therefore the unconfoundedness or MAR assumptions may not hold true conditional on available *X*. In this case, estimates of measurement or selection effects would be still biased.

1.5 Contributions and Outline of this Dissertation

In the absence of true scores and given the confounding problem in assessing measurement equivalence, it is very difficult to evaluate different mixed-mode strategies in survey practice at the design stage. This dissertation develops and applies methodology for evaluating research objectives 1 to 3 in the absence of true scores and while controlling for the confounding problem. All methodological work is based on experimental data collected in a multi-mode experiment at Statistics Netherlands in 2011. These data are partly used for empirical studies and partly for illustrating the developed methodology. An overview on the experiment is provided in the next section. Afterwards, the content of the remaining chapters is outlined.

1.5.1 The MEPS experiment

As part of this dissertation, a large scale mode experiment was designed and executed in collaboration with Statistics Netherlands for the case of the Dutch Crime Victimization Survey (CVS), called the MEPS experiment (in Dutch: *mode effecten in persoons enquêtes* / *mode effects in social surveys*). As noted in the previous chapters, the regular CVS, a yearly national survey, has undergone several redesign stages in the past decades. In 2008, a sequential mixed-mode design was introduced (Figure 1.1, middle), which showed large total differences in estimates taken from different mode-specific response groups. Time-series of victimization and further national CVS statistics became relatively unstable in the course of the subsequent years. The primary objective of the MEPS experiment was to assess mode differences in selection error and measurement error (measurement equivalence), and derive conclusions about a more optimal mixed-mode data collection design for the CVS based on the findings.

The experiment consisted of a two-wave design (Figure 1.2). The first wave entailed a split ballot design that assigned a sample of 8,800 individuals drawn from the person register of The Netherlands to one of four modes (2,200 individuals each): CAPI, CATI, mail (paper-SAQ), and web (web-SAQ). In the second wave, all units were approached again, but now most units received an in-person follow-up (CAPI). A small share of the second wave was approached in telephone (CATI) to reduce the costs of the second wave. In the second wave, a set of CVS questions was repeatedly measured and, additionally, further information, such as attitudes on surveys and on the response process at the first wave were collected. On a conceptual level, it was planned to use the second wave data as additional auxiliary information in estimating the confounded measurement and selection effects between the single-mode designs.

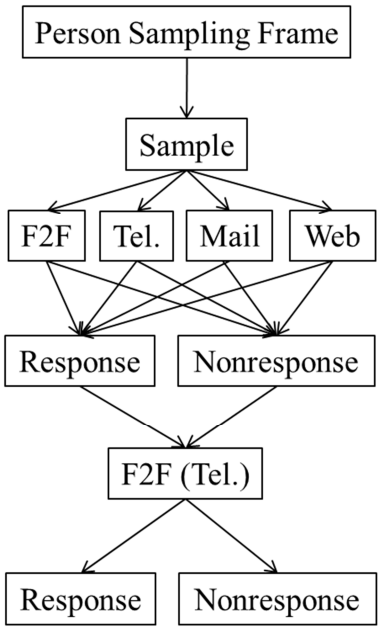


Figure 1.2: Illustration of the MEPS experiment in the Crime Victimization Survey

Whereas the primary objective of the mixed-mode experiment was a comparison of single-mode surveys, the design also entailed a sequential mixed-mode (Type III) element. Nonrespondents (and non-covered units) in each of the four mode-specific samples were approached in either face-to-face or telephone. This contact approach resembled a 'real-world' sequential design to great extents. For example, the fieldwork period between waves was four to six weeks, which is comparable to real sequential mixed-mode surveys. Also the contact attempt in face-to-face or telephone follow-ups of nonrespondents was similar to sequential mixed-mode surveys at Statistics Netherlands. Finally, following up nonrespondents in inexpensive modes, such as web or mail, by more expensive modes is a common design option in survey practice (cf. section 1.2.3). This extension of the MEPS design allowed assessing three sequential mixed-mode surveys: telephone, mail, and web, followed by F2F (the telephone follow-up sample was considered too small for substantial analyses of similar type).

The following section explains how these data were applied in addressing the general research objectives 1 to 3 in mixed-mode practice. Each of chapters 2, 3, and 5 entails an empirical results section using data from the MEPS experiment. Chapter 4 applies the data for illustrating a more general set of methods. Each chapter introduces the part of the MEPS data relevant for the particular problem addressed. Further

information on the experimental design is provided by Buelens, Van der Laan, Schouten, Van den Braken, and Klausch (2012) and Klausch, Hox, and Schouten (2013a).

1.5.2 Outline of chapters two to six

In chapter 2, research objective 1 is addressed, evaluating the selection error of single- and mixed-mode surveys. The MEPS experiment allows answering two research questions within this general objective. First, the chapter evaluates how large selection error of the single-mode surveys in the CVS experiment is (face-to-face, telephone, mail, and web) and how strong mode-differences in selection error (selection effects) are. The size of these effects provides evidence on the importance of mode choice for impacting the selective process of a survey design. Second, the question which impact a sequential face-to-face follow-up has on the absolute size of selection error of the single-mode surveys in the CVS receives attention. The size of these effects has implications for the necessity of mixed-mode designs to reduce selection error.

Methodologically, selection error is studied based on summary statistics on variable level and across variables. For this purpose, benchmark data is made available. Two types of data are used for this purpose. First, a series of socio-demographic characteristics is available on the sampling frame level from the Dutch national register. Second, the second wave of the MEPS experiment is used as a re-interview design (i.e., face-to-face). Re-interview designs are classical approaches to studying survey error in the absence of true scores (Biemer & Lyberg, 2003, pp. 283–301). For example, nonresponse follow-up surveys using the call-back approach assess survey nonresponse error by comparing response means of initial and re-interviewed respondents (Dunkelberg & Day, 1973; Elliott, Little, & Lewitzky, 2000). However, no empirical study so far has used a re-interview approach in a mixed-mode context. The repeated measure in the face-to-face re-interview solves the confounding problem of measurements and selection error at the first wave (cf. section 1.4.2), since regardless of mode assigned at the first wave, measurements in a single mode are obtained at the second wave. Based on the empirical findings, it is discussed how important mode choice is for survey selection error in the presence of clear differences of response rates between designs in the MEPS.

In chapter 3, research objective 2 is addressed. It is suggested to evaluate measurement equivalence between survey modes empirically using a multiple-group confirmatory factor analysis (MCFA) model. This model applies particularly well to evaluating measurement error components of attitudinal rating scale questions. Each measurement model consists out of three components: scale parameters describing the relation of observed and a latent (unobserved) true score, random measurement error (question reliability), and a systematic error component attributed to the method (method bias and variance). The latter type of measurement error is not question specific, reflecting that the method (e.g., mode or scale-type) may contribute equivalent bias and variance across sets of questions in a scale. These parameters allow evaluating three types of measurement effects: question-specific effects due to differences in scale parameters (effect on systematic measurement error bias of particular questions), effects on random measurement error (question reliability), and effects on method-level measurement bias and variance. The distinction of question-level and method-level error components allows generalizing conclusions of mode impact in the measurement process beyond question-specific content. To address the confounding problem of measurement and selection effects, a method called 'propensity score weighting' of the four single-mode response samples in the CVS experiment to the full sample is applied (cf. Figure 1.2). For this purpose socio-demographic register information is used as auxiliary information. In doing so, measurement equivalence on population level (marginal measurement effects) is assessed (cf. section 1.3.2). Based on the findings it is discussed, which modes may be suitable for combination in sequential or concurrent mixed-mode designs.

In chapter 4, an alternative look at the evaluation of measurement equivalence is taken. The chapter begins by the observation that the available auxiliary data in single-mode comparisons, but also in mixed-mode data collection designs (sequential / concurrent), is often insufficient to plausibly ignore the selection mechanism between modes (cf. section 1.4.2). Therefore, in practice, measurement effect estimates may still be biased after adjustment for commonly available data, such as socio-demographics. To address this deficiency, the central objective of this chapter is the development of a framework, in which more auxiliary variables can be made available for making the unconfoundedness assumption more plausible. Such a framework is offered by 'within-

subject' designs, in which each unit is observed under two different modes at two subsequent points in time. Due to the repeated measurement, additional information is made available that can be used in adjusting for selectivity between modes under more plausible assumptions than commonly possible in practice. It is discussed how different conditional measurement effects can be estimated from within-subject data and which assumptions different estimation approaches make. The MEPS experiment is used for illustrative purposes, given that it made available within-subject data. It is also explained how the conditional measurement effects can be used to take decisions about mixed-mode data collection designs of types III, IV, and V in real-world decision situations.

In chapter 5, research objective 3 is addressed for the case of the MEPS data. As noted above, the design made available three sequential mixed-mode surveys: telephone, mail, and web followed by face-to-face. The objective of this project is to evaluate which alternative benchmark data may be used in the absence of true scores. One important option is to define measurements of benchmark surveys, also called preferred mode or golden standard, as optimal (cf. section 1.4.2). These scores become a substitute to the latent true scores when assessing total survey error (for example, face-to-face measurement of CVS questions). If measurements from a single-mode survey are used for this purpose, the resulting benchmark is called a single-mode benchmark (e.g., face-to-face survey). Besides the measurement error bias of a single mode, the single-mode benchmark is also characterized by the selection mechanism of the single-mode survey (and hence also the selection error bias). For this reason, the single-mode benchmark entails two sub-types of benchmark assumptions, the measurement benchmark and the 'selection benchmark'. The idea of a selection benchmark is based on the notion that the response propensities in the selection benchmark mode are more balanced in the population than in the comparison designs. Subsequently, it is noted that a single-mode benchmark may often offer a sub-optimal error balance in theory. For example, a face-to-face benchmark survey may suffer from social desirability bias, but it may lead to very representative samples (small selection error) due to its typically high response rates (or, more generally, a 'desirable' selection mechanism). For this reason, additionally, the so-called 'hybrid-mode benchmark' is introduced, a hybrid form of measurements from one mode (e.g., web) and the response mechanism from a

second mode (e.g., face-to-face). The hybrid-mode benchmark is chosen to optimize a benchmark estimate with respect to measurement and selection. Based on empirical estimates of total survey error bias, measurement error bias, and selection error bias against the two benchmarks, conclusions about the usefulness of sequential mixed-mode data collection strategies in the CVS are drawn.

In chapter 6, the methodological developments are discussed and set into more general context. Furthermore, an outlook is provided on the most important next steps in methodological research about mixed-mode data collection designs.

Chapters two to five in this book were written as independent publications either published with or under review at international journals. For this reason, a certain redundancy in the introduction of the MEPS experiment is present across these chapters. However, each chapter highlights only those elements of the research design that are crucial for the problem at hand. An advantage of this procedure is that all chapters can be read and understood independently from each other.

2 Evaluating Selection Error in Single- and Mixed-Mode Surveys using a Re-interview Approach [1]

The extent of expected coverage and nonresponse error is an important quality criterion of a survey design (Biemer & Lyberg, 2003, p. 39; Groves, 1989, pp. 15–18). The 'survey mode' might play a crucial role in determining the size of these errors. Clearly, modes determine applicable sampling frames (e.g., to telephone households) and thus impact population coverage of a survey. Furthermore, modes are related to response rates (RRs), which might be indicative of the extent of nonresponse error evoked by a survey design. Meta analyses have reported that face-to-face modes (F2F) yield higher RRs than telephone or mail modes (Hox & De Leeuw, 1994) and that web modes evoke lower RRs than the other major modes (Manfreda et al., 2008; Shih & Fan, 2008). Therefore, F2F might cause smaller nonresponse error than telephone and mail, and nonresponse error in web might even be stronger.

Unfortunately, F2F surveys commonly incur the greatest costs during data collection. So-called 'sequential mixed-mode surveys' can cope with this problem by offering inexpensive modes first (e.g., web, mail) and later follow-up only the nonrespondents by a more expensive mode, such as F2F (Biemer & Lyberg, 2003, p. 106; De Leeuw, Dillman, et al., 2008; De Leeuw & Hox, 2011; De Leeuw, 2005; Dillman & Christian, 2005; Dillman, Smyth, et al., 2009, pp. 300–310). In doing so, high RRs (e.g., at the level of F2F) can typically be preserved. A prominent example is

[1] This chapter is conditionally accepted at *Journal of the Royal Statistical Society: Series A* as Klausch, T., Hox, J., & Schouten, B., Selection Error in Single and Mixed-Mode Surveys of the Dutch General Population.

the American Community Survey, in which mail nonrespondents are followed up by telephone and F2F. This strategy, though participation is mandatory, strongly increases RRs by the mode switches to over 90% (US Census Bureau, 2010). Several further studies also report response increases, when mail or web modes are combined with telephone in sequential designs (Dillman, Phelps, et al., 2009; Eva et al., 2010; Fowler et al., 2002; Greene, Speizer, & Wiitala, 2008; Link & Mokdad, 2006; cf. Millar & Dillman, 2011 for a web-mail comparison). For example, Link and Mokdad (2006) report up to 39 %-points additional response in telephone follow-ups of web nonrespondents. Similar results (35 %-points) are reported by Dillman, Phelps *et al.* (2009; cf. De Leeuw, 2005 and Lynn, 2013 for further examples). These effects might imply that mixed-mode surveys can also mitigate coverage and nonresponse error vis-à-vis single-mode designs with lower RRs. Keeping these errors low is referred to as a key advantage of mixed-mode surveys in many of these studies.

However, using response and coverage rates as indicators for the risk of nonresponse and coverage error is based on the assumption that the quantities are inversely related. This conjecture appears contestable on theoretical and empirical grounds (Wagner, 2012). Statistically, RRs are not directly related to the size of nonresponse bias (see, for example, Bethlehem, 1988). Empirically, it has been shown by extensive meta-analysis that nonresponse bias and RRs are only weakly associated (Groves & Peytcheva, 2008; Groves, 2006). Basing design decisions on response and coverage rates, therefore, poses a potential fallacy. However, only few studies have assessed both errors empirically for multiple single- and mixed-mode surveys.

The present study addresses this issue by a large-scale (n=8,800) mode experiment within the Dutch Crime Victimization Survey (CVS). We consider nonresponse and coverage errors as a compound, because practitioners generally are concerned most about the 'net effect' of both error sources when taking design decisions. This compound is referred to as the 'selection error' of a design (also known as non-observation error, Groves, 1989, pp. 15–18). In doing so, we address two research questions:

- Do mode-specific differences in RRs reflect differences in selection error? (RQ1)

- Are RR increases of sequential mixed-mode designs indicative of a reduction in selection error? (RQ2)

In particular, we first study selection error of the four major contemporary modes (F2F, telephone, mail, and web) and, subsequently, illustrate the impact of three sequential mixed-mode designs on selection error of the single-mode designs (telephone-F2F, mail-F2F, and web-F2F). Empirical evidence about selection error in these designs can support survey practitioners in re-evaluating the importance of response and coverage rates when designing single- and mixed-mode surveys.

Two types of benchmark variables are applied in the assessment of selection error: socio-demographics and target variables from the CVS. Unfortunately, it is a common problem in mode experiments that mode differences in measurement error are confounded with mode differences in selection error (Vannieuwenhuyze & Loosveldt, 2013). Therefore, it is an important aspect of this study that socio-demographics were available from an external source: the national register, for which measurement error is known to be small (see Bakker, 2012 for a validation study of the Dutch register). Furthermore, the present study is the first to evaluate design effects on selection error also for survey target variables. In doing so, a specific two-wave experimental design allowed studying selection error independently of measurement error. Details on this problem are discussed in section 2.1 and the two-wave design is explained in section 2.2. We present the analytical approach in section 2.3, results in section 2.4 and conclude by section 2.5.

2.1 Prior Research on Selection Error in Single and Mixed-Mode Surveys

The choice of survey mode can affect selection error at different stages of the fieldwork process, in particular when defining sampling frames and when contacting and convincing individuals for participation (Groves et al., 2010, pp. 192–201). Mixed-mode designs might reduce selection error, furthermore, because advantages from different stages of this process are combined, for example by offering potential nonrespondents an alternative response channel (De Leeuw, 2005; Dillman, Smyth, et al., 2009, pp. 306–310; Groves et al., 2010, pp. 175–177). This section considers the

prior empirical research on differences in selection error across single- and mixed-mode surveys. Consequently, the improvements of the present study over previous scholarly work are pointed out in detail.

2.1.1 Empirical evidence for mode differences in selection error

It is a widely recognized problem of comparative mode research that response distributions of survey target variables confound mode differences in selection error with mode differences in measurement error, if modes are administered to different samples (Jäckle, Roberts, & Lynn, 2010; Vannieuwenhuyze & Loosveldt, 2013). For this reason, survey variables can only be used for studying mode differences in selection error, if it can be assumed that measurement error is absent or equal across modes. This condition is likely not to hold for many attitudinal or factual questions.

Prior studies mainly assessed selection error on socio-demographic variables. In the past years pairwise comparisons involving web and either mail or telephone have prevailed, as interest was set on assessing 'the new' mode versus 'the older' modes (Bälter, Bälter, Fondell, & Lagerros, 2005; Dillman, Phelps, et al., 2009; Link & Mokdad, 2005; T. I. Miller, Kobayashi, Caldwell, Thurston, & Collett, 2002). There has been great variance in the quality of data across studies, unfortunately. Often, special interest groups instead of a general population could be studied, for example students or employees (Kaplowitz, Hadlock, & Levine, 2004; Kwak & Radler, 2002; Sax, Gilmartin, & Bryant, 2003). On other occasions, pre-selection of the general population limited generalisability of results, for example by sampling from web panels (Berrens et al., 2003; Braunsberger, Wybenga, & Gates, 2007; Chang & Krosnick, 2009), by RDD screeners (Fricker, Galesic, Tourangeau, & Yan, 2005), or due to fieldwork errors (Dillman, Phelps, et al., 2009). Consequently, the results of these studies are rather inconsistent possibly owing to the differences in data quality or populations. Perhaps the most consistent finding is that web respondents are somewhat higher educated and more affluent reflecting typical characteristics of web coverage error (Berrens et al., 2003; Couper et al., 2007; Couper, 2000; Kwak & Radler, 2002).

The aforementioned studies often implicitly assume that socio-demographic questions are answered identically in all modes and therefore the analyses do not

confound measurement error differences with selection error differences. This belief, however, has not been tested and may be wrong. The assumption is only valid with certainty, if socio-demographics are available from external benchmarks, such as registers (Groves, 2006). When a register is the source of auxiliary information, measurement error is always distributed equally across modes provided that random sampling is used. Register variables, however, often were not available in prior studies and the socio-demographic characteristics had to be surveyed within the mode experiments. For the same reason, the absolute size of selection error rarely could be assessed, because only relative comparisons between modes are possible in the absence of an external benchmark. An absolute assessment is however required to answer the type of RQs addressed in the present study.

2.1.2 Impact of mixed-mode designs on selection error

Mixed-mode surveys often apply a combination of self- and interviewer administered modes in sequential designs, which are known to increase RRs consistently (cf. introduction to this chapter). There is, however, scarce knowledge about the effects on selection error (De Leeuw, 2005). Dillman, Phelps *et al.* (2009) assessed mail-telephone and web-telephone combinations and found only marginal or no change in selection error across multiple socio-demographic variables. Similar results are reported by Schouten, Cobben, and Bethlehem (2009). However, Fowler *et al.* (2002) report on a study in which single-mode mail was biased for gender and age distributions, which could be mitigated by approaching mail nonrespondents by telephone. Voogt and Saris (2005) report on a complex sequential mixed-mode design, which could reduce the nonresponse bias on estimates of voting turnout. Also, Link and Mokdad (2006), who compared a web and mail survey with telephone follow-up to regular telephone response, found with respect to race, income and education, less selection bias in the joint samples, but age distributions were more biased than in the telephone mode alone. These results imply that findings are again inconsistent across variables and studies. It is possible that mixed-mode surveys are capable of reducing error on some variables while increasing it on others. In this case, it would be desirable to know, however, whether the reduction in error outweighs the increase, which to date has not been assessed.

2.1.3 Improvements of the present study

The present study aims to cope with some of the deficits of this earlier empirical work. Firstly, we improve on data quality. We present a full comparison of the four major contemporary modes for the same survey based on a large probability person sample from the general population (F2F, telephone, mail, and web). Furthermore, our research design allowed assessing how selection error of the single-mode designs is impacted by sequential mixed-mode extensions, in which the temporary nonrespondents from telephone, mail, and web are followed up by a second mode (F2F). This variety of possible comparisons and statistical power is new.

Secondly, we improve on benchmark variables. Earlier work often considered only relative differences in selection error on socio-demographics, lacking an external benchmark. In the present study, socio-demographics of all subjects were available from a high quality national register. Whereas socio-demographic sample profiles are important, it is, however, seldom the case that error on socio-demographics is also indicative of error on survey target variables, especially if target variables and socio-demographics are weakly related, a situation often encountered in practice. Therefore, the present study extends conclusions to survey target variables. To achieve this goal, we had the unique opportunity to measure key target variables using a re-interview approach in a single mode (face-to-face), as explained in detail in section 2.2. We applied a nonresponse correction to extrapolate conclusions from the re-interview to the population (multiple imputations using register information, cf. section 2.3). Generalizations to the population depend on the efficacy of this procedure, but are at least valid for the respondents to the face-to-face survey.

Thirdly, we improve on the statistical methodology. Prior research has considered selection error in isolated, univariate analyses of different benchmark variables. Drawing conclusions about the impact of mode is difficult, however, if findings differ by variables, as we observed for some mixed-mode studies that found increases in error on some variables and decreases on others. To answer the RQs it is then more important, whether single- and mixed-mode designs still cause 'systematic differences' in selection error, even though there might be variable specific variations. Conclusions, such as 'F2F evokes on average less selection error than other designs', are only

possible, if information about selection error can be combined across variables. Section 2.3 provides details on the options to study such effects.

2.2 Research Design

We conducted a mode experiment with two waves within the national Crime Victimization Survey (CVS) carried out by Statistics Netherlands. In the first wave, a probability person sample of 8,800 individuals was drawn from the national register, a list of all individuals living in The Netherlands that can be considered free of coverage error. The sample was randomly split between four modes: F2F, telephone, mail, and web (2,200 individuals each). This design allowed estimating and comparing selection error for a set of benchmark variables across modes (RQ1). The benchmark measures applied in the present study were partly available from the register and partly collected during the second wave, as described in detail next. Afterwards, we provide details on the mixed-mode data used for RQ2 and provide details on fieldwork procedures.

2.2.1 Measures for studying selection error

We intended to study selection error on two types of variables: socio-demographics and target variables from the CVS. A set of six socio-demographic variables was available from the national register: sex, age, income, ethnicity, marital status, and household size. We additionally included two geographical indicators (degree of urbanization and living in one of the three large cities of The Netherlands). These variables represent a high quality (i.e., small measurement error, cf. Bakker, 2012), exogenous benchmark that is available for all units regardless of being nonrespondents or respondents in any mode. Furthermore, since these variables were available from an external register, differences in measurement error on socio-demographics across modes can be excluded for the present study.

To cope with the difficulty of confounded selection and measurement error on target variables observed during the first wave, we administered another survey to the same sample about 4-6 weeks later. This ‘second wave’ approached every individual again using only the F2F mode, regardless of mode assigned at wave 1 and being respondent or nonrespondent. A core set of questions from the CVS was repeated in this

survey. These second wave F2F measurements could now be regarded as a benchmark exhibiting equivalent measurement error regardless of mode assigned at wave 1 (i.e., measurement error of F2F). The wave 2 data could be compared across wave 1 respondents and nonrespondents to assess selection error. These evaluations were not confounded with measurement error anymore, since all variables were measured by F2F in wave 2.

The set of repeated CVS variables was rich, including 22 attitudinal and factual questions. Attitudinal questions were asked about the social quality, security and problems of the neighbourhood. Factual questions concerned the time of past police contact and past victimization to different forms of crimes.

However, the second wave, like any survey, suffered from unit nonresponse (48.6%). Since we needed the full second wave sample as a benchmark for valid inference about the population, the missing data problem had to be adequately addressed. We used multiple imputations of unit nonresponse for this purpose, as explained in detail in section 2.3 (Rubin, 1987, pp. 15–17; Schafer & Graham, 2002; van Buuren, 2012, pp. 25–49). The socio-demographics from the register could be used to build the imputation models, because they were available for all units in the sampling frame.

2.2.2 Mixed-mode data collection

After having assessed the difference in selection error between modes at the first wave, we considered the change brought to selection error by approaching wave 1 nonrespondents (including non-covered units) by F2F used during the second wave (RQ2). The two-wave experimental design used to collect F2F measurements of target variables thus additionally made available data that strongly resembled three sequential mixed-mode surveys, in particular: telephone – F2F, mail – F2F, and web – F2F. These designs are typical for survey practice, because a more expensive mode (F2F) with a higher response potential is used to follow-up nonrespondents to cheaper modes with lower response potential (De Leeuw, 2005).

2.2.3 Fieldwork

Fieldwork took place in the period of April to June 2011. First, we posted personalized invitation letters to all mail addresses. These letters differed only by information on how to participate in the survey: the mail condition contained a paper questionnaire, letters in the web condition mentioned a hyperlink to the online survey, and letters in telephone and F2F informed individuals about forthcoming interviewer contact. Incentives were not offered in any of the conditions. Individuals were not informed about wave 2 at this point to avoid biasing response samples at wave 1 for respondents who preferred a F2F mode.

During the subsequent fieldwork of four weeks used for wave 1, individuals in the two self-administered modes received up to two mailed reminders before being classified as wave 1 nonrespondents. In the telephone condition, multiple call attempts were made to individuals with a known landline telephone number. 28.5% of all persons allocated to the telephone condition were not covered by lists available to Statistics Netherlands, however (or numbers turned out incorrect during fieldwork). Non-covered persons were classified as telephone nonrespondents directly. In the F2F condition, finally, the interviewers tried to establish contact on location for a maximum of six times.

Table 2.1 provides information on the response turnout of the two-wave design. The first column shows response rates (RRs) to the first wave, the second column RRs to the second wave and the third column the mixed-mode RRs. Here the term RR refers to the proportion of respondents out of all eligible units covered by the person sampling frame including units not covered by telephone (also discussed as 'realization rate' in Skalland, 2011; this definition deviates from the AAPOR RR1 standard, which would exclude non-telephone households in the telephone sample before computing telephone response rates). The first column, 'Wave 1 Sample', shows RRs by mode conditions. Originally, 2,200 individuals were assigned to each condition, but frame errors (e.g., due to relocation) reduced these numbers slightly (eligible units indicated in the parentheses next to the RRs). In F2F, 64.3% of eligible persons responded, 48.2% in telephone, 49.8% in mail, and only 28.7% in web. This order of RRs matches experience of earlier studies stressing the point that highest response can be expected

from F2F and lowest from web (cf. introduction to this chapter). The results section assesses, whether these differences in RRs also imply mode differences in selection error addressing RQ1 (the AAPOR RR1 for the telephone sample was 67.7%, which excludes non-telephone households before calculating the RR1; although this RR was substantially higher, our RQs consider the combined impact of nonresponse and non-coverage and so do the RRs reported in Table 2.1).

The second wave followed four to six weeks after the first. Due to cost reasons, a smaller sample of wave 1 units was randomly selected (approx. 80%, n=6,803). Telephone households, furthermore, were oversampled for reasons unrelated to the present study. All subsequent analyses involving wave 2 used design weights to adjust for this overrepresentation. The weighted RRs at wave 1 in this sub-sample closely followed the RR indicated in the first column of Table 2.1 for the full sample[2].

Table 2.1: Response rates of wave 1 and 2 by mode of administration, and the change induced by a F2F follow-up to nonrespondents at wave 1 (mixed-mode response)

Mode assigned at wave 1	RR Wave 1 % (n)	RR Wave 2 % (n)	RR Mixed-Mode % (n)	Δ RR
F2F	64.3 (2,081)	53.2 (1,639)	(71.1)[a] (1,639)	(+6.7)[a]
Telephone	48.2 (2,062)	50.1 (1,658)	65.2 (1,658)	+16.6
Mail	49.8 (2,182)	51.9 (1,760)	67.3 (1,760)	+18.0
Web	28.7 (2,199)	50.5 (1,746)	59.6 (1,746)	+30.7
Total	47.5 (8,524)	51.4 (6,803)	65.7 (6,803)	+18.1

a. For completeness the increase in response rates for the wave 1 F2F- wave 2 F2F combination is reported. In this case the wave 2 F2F survey represents a simple nonresponse follow-up survey for first wave F2F nonrespondents and not a mixed-mode design.

Contrary to wave 1, no separate invitation letters were sent at the outset of the second wave. Instead, F2F interviewers were instructed to explain to individuals the need for (repeated) participation in the CVS. This procedure was chosen in order to keep the perceived survey burden for respondents at the first wave low, which could have been increased significantly by communicating the second wave as a separate survey.

[2] It should be noted that 500 respondents from the first wave of interviews were classified as not re-approachable in the second wave during fieldwork of the first wave (e.g., due to long-term illness or hard refusal). A stratified sample (by mode and telephone coverage) of 399 respondents was assigned to the sample of the second wave and treated as wave 2 nonrespondents, but this group was not re-approached in the fieldwork of wave 2.

The second column shows RRs at the second wave. 3,498 individuals responded, which represents 51.4% of all eligible persons (n=6,803). Although the response at wave 2 was overall lower than for F2F at wave 1 (64.3%), the variation of wave 2 RR across modes assigned at wave 1 was not significant ($\chi^2 = 4.07$, *df* = 3, n.s.). Hence we may assume that the unwillingness for repeated participation at wave 2 was distributed evenly across wave 1 modes. There was also no systematic difference in socio-demographic indicators between the F2F response sample at wave 1 and wave 2 (Klausch et al., 2013a; Schouten, van den Brakel, Buelens, van der Laan, & Klausch, 2013). Both conjectures support the claim that the full second wave was very similar to a standard F2F survey.

Finally, the third and the fourth column of Table 2.1 illustrate the change in RRs caused by adding wave 2 F2F respondents who were nonrespondents or not covered at wave 1 to the mode-specific response samples (i.e., mixed-mode response, RQ2). All mixed-mode designs yielded a strong response increase (Δ RR). This change was particularly pronounced for the web mode (+30.7). About half of the mixed-mode web sample was represented by second wave F2F respondents. In telephone and mail this effect was still strong (+16.6 / +18.0). The power of the mode switch is stressed by comparing these figures to the F2F condition, where response was only marginally increased (+6.7%). These RR shifts reflect experiences reported in the literature (cf. introduction to this chapter) and can be considered typical for sequential mixed-mode surveys of this type. In the analyses we assess the implications for selection error (RQ2). However, before we present these results, the next section describes our statistical approach.

2.3 Statistical Methodology

In this section, we describe how selection error was studied statistically. Answering the RQs required two steps. First, we measured selection error in each mode and compared it across the single-mode designs. Second, we assessed the impact of the mixed-mode extension on the selection error of the single-mode design (i.e. we considered whether it was decreased, increased, or stayed equal). In doing so it was relevant to specify what we mean by selection error. We see three relevant ways of specifying selection error:

- Absolute selection error per benchmark variable,
- Relative selection error per benchmark variable, and
- Absolute selection error over sets of benchmark variables.

Absolute and relative selection errors are specific to a given variable. Analyses based on these statistics thus may lead to variable-dependent conclusions about the RQs. The third approach therefore tries to generalize conclusions on absolute selection error across sets of variables to answer the RQs (cf. section 2.1.3). After introducing these perspectives, we provide details on how a multiple imputation technique was used to allow inference to the population for target variables measured in wave 2.

2.3.1 Absolute selection error

Suppose a survey design using mode *M* evokes selection error on a given variable *Y*. Then the distribution of *Y* differs between response and nonresponse groups identified by response mechanism *S*. We can write the conditional distribution of *Y* as $P(Y|S,M)$. Selection error is absent in a specific mode $M=i$, when *Y* is conditionally independent of *S*:

$$P(Y|S,M=i) = P(Y|M=i). \tag{2.1}$$

Our benchmark variables *Y* (e.g., socio-demographics) were available on discrete measurement level with differing numbers of categories. In principle, definition (1) could be applied to each category-specific estimate of selection error by assessing its size and comparing it across designs. However, this procedure would lead to a large number of estimates increasing with the number of categories, variables and design comparisons.

We therefore used a summary statistic for the strength of absolute deviations from independence (1) in mode $M=i$ across all categories of *Y* on the variable level, called the absolute selection error. Our analyses are based on Pearson's chi-square statistic which sums the squared category-specific deviations from independence (i.e., absence of selection error) across cells of the contingency table of *Y* and *S*. However, chi-square has to be scaled in order to be comparable across tables. A useful measure for this purpose is Cramér's *V*, scaling chi-square to the interval of 0 and 1. *V* has an effect size

interpretation (e.g., .10-.30 small selection error, .30-.50 medium, and .50-1.00 large). *V*, however, is by far not the only measure of strength of association that could be applied. Alternatives are represented by, for example, the dissimilarity index (Agresti, 2002, pp. 329–330) or Cohen's *w* (Cohen, 1977, pp. 222–224). All analyses were also executed for the alternative measures with equivalent results. We present results on *V*, because of the clear effect size interpretation of the index.

Furthermore, testing independence based on Pearson's chi-square or likelihood-ratio statistics indicates the statistical significance of associations. However, in particular in large samples even small deviations from independence can lead to significant tests. In these cases, effect size measures such as *V* are more informative about the strength of deviation from independence (i.e., absolute selection error).

To answer RQ1 for a survey variable *Y*, we estimate *V* for each mode and compare it across modes. To answer RQ2, the change in *V* by the F2F follow-up is assessed.

2.3.2 Relative selection error

It is a disadvantage of measures of absolute selection error, that the direction (sign) of error is ignored across categories. Equivalent Cramér's *V* indices, therefore, do neither suggest that selection error is present on the same categories nor that error has the same sign. Instead *V* merely indicates that across categories absolute deviations from expected frequencies are equal. For this reason it is additionally informative to consider relative selection error, which is based on the following independence relation:

$$P(Y|S = 1, M) = P(Y|S = 1). \tag{2.2}$$

The left side of the equation considers the response distribution of *Y* conditional on mode. If the response distribution depends on mode *M*, there may be both positive and negative deviations from the unconditional response distribution $P(Y|S = 1)$. The conditional response distribution can be thought of as a contingency table of *M* and *Y* with frequency counts from the response sample only. For example, consider a table of categorized income bounds against the four modes assessed in this study. Each cell contains the number of respondents in a particular mode and an income group. The level of dependence of mode and income consequently indicates the strength of relative

differences in selection error, which again can be measured by Cramér's V for the contingency table.

However, absolute differences against the population cannot be seen from relative selection error (i.e. definition 1 is needed). For example, even if equation (2.2) holds and there are no relative differences in selection error, there can still be selection error against the population if $P(Y|S = 1) \neq P(Y)$. Since this information is required to answer the RQs (e.g., for RQ2 we need to know whether selection error was increased or decreased by the mixed-mode follow-up, which cannot be seen from relative selection error), we focused on design differences in absolute selection error in all analyses and considered relative selection error as secondary information.

2.3.3 Absolute selection error for sets of benchmark variables

With this level of analysis, we want to extract systematic impact of mode on multiple variables. The benchmark variables may show a diffuse picture in their absolute selection error over designs due to factors related to variable content. In this case, however, it becomes difficult to attribute the differences in selection error to the mode. Assessing systematic differences in absolute selection error aims to isolate the joint variance in absolute selection error across variables, which is due to the survey design and not variable content. In the literature, for example, there were inconsistent findings across variables when assessing the impact of mixed-mode designs on selection error (cf. section 2.1.3). In such a situation it is helpful to extend variable-level analyses to a summarizing view on selection error across variables. For this purpose, we chose two different strategies.

First, we compared the distribution of all variable-specific estimates of absolute selection error across the single-mode designs addressing RQ1 (i.e., all Cramér's V estimates from definition 1). The variation of V statistics across variables and their central tendency allows conclusions about which design 'on average' causes less absolute selection error and how much variance can be attributed to variable content. To assess RQ2, the impact of the mixed-mode follow up on the single-mode distribution of V was considered.

Second, we applied an index which summarizes deviations from independence in (1) across multiple variables simultaneously: the 'representativeness (R-) indicator' (Schouten et al., 2009; Schouten, Shlomo, & Skinner, 2011; Shlomo, Skinner, & Schouten, 2012). R-indicators are a new class of quality indicators defined on the variance of response propensity represented by $P(S = 1|Y, M = i)$. If selection error is absent, (1), it holds that

$$P(S = 1|Y, M = i) = P(S = 1|M = i). \quad (2.3)$$

In other words, response propensities then are constant across *Y*. Therefore, deviations from constant response propensities, measured by their variance, are informative about the extent of selection error on *Y*. Schouten *et al.* (2009) have shown that when estimating $P(S = 1|Y, M = i)$ for a set of multiple benchmark variables *Y* by means of a logit model, the variance of response propensities indicates the extent of selection error on all underlying variables. The *R*-indicator (in short, '*R'*) is then defined as

$$R_{M=i} = 1 - 2sd[P(S = 1|Y, M = i)], \quad (2.4)$$

where '*sd*' denotes standard deviation. *R* varies between 0 and 1, where values close to 1 indicate absence of selection error ('representative response'). To assess RQ1, *R* can then be compared across the single-mode designs to assess systematic differences in selection error (i.e., for all benchmark variables *Y* simultaneously). To assess RQ2, the change in *R* by the mixed-mode follow-up is considered.

As the response propensities are estimated from a logit model, *R*-indicators additionally control for the multi-colinearity in the benchmark variables, isolating the covariance with *S* that is uniquely caused by an underlying selection mechanism. Therefore, analyses of *R* and mean Cramér's *V* might differ when considering multiple *Y*. For single variables *Y*, however, Cramér's *V* and *R* are asymptotically equivalent in large samples except for a scaling constant (Schouten et al., 2009) (Schouten *et al.*, 2009). *R*-indicators, however, still lack a clear effect size interpretation like *V*, but feature the advantage to summarize the variance of response propensities over multiple variables into a single number. Variance estimation for R-indicators is discussed in Shlomo *et al.* (2012).

2.3.4 Inference to the population for wave 2

As pointed out in section 2.2, the second wave suffered from unit nonresponse (48.6%). We imputed the missing cases by an algorithm called 'multiple imputations by chained equations (MICE)' using socio-demographic frame information as auxiliary information (van Buuren & Groothuis-Oudshoorn, 2011; van Buuren, 2012, pp. 20–49). This method assumes that units were missing at random (MAR) in wave 2 given register information (Little & Rubin, 2002, pp. 12–19). It is important to note that inference to the population is only valid given this assumption.

Ten multiply imputed data sets were analysed separately and then pooled by taking the mean of V and R-indicators across the imputed data sets (Rubin, 1987, p. 90; Schafer & Graham, 2002). P-values for tests of independence in multiply imputed contingency tables were computed following a procedure for the pooling of Chi-squared distributed statistics described by Li and colleagues (Li, Meng, Raghunathan, & Rubin, 1991; Schafer, 1997, pp. 115–116; van Buuren, 2012, p. 159). The standard error of R-indicators is estimated by pooling the between- and within-imputation variances of R using Rubin's rules (Rubin, 1987, p. 90).

2.4 Results

We present analyses of selection error on eight register variables and 22 CVS target variables. First we consider RQ1, estimating absolute and relative selection error in each of the four single-mode designs per benchmark variable and across sets of variables. Subsequently, RQ2 is assessed by illustrating the impact of the three mixed-mode designs on absolute selection error of the single-mode designs.

2.4.1 Single-mode differences in absolute selection error (RQ1)

RQ1 asked, whether the order of RRs reflected the extent of selection error of the four single-mode surveys: F2F, telephone, mail, and web. We found largest response in F2F, followed by mail and telephone (Table 2.1). Response was lowest in web. Table 2.2 provides Cramér's V statistics for eight register variables. It is instructive to consider first relative selection error (formula (2.2)). Here, V is estimated for each contingency

table of a benchmark variable and mode conditional on response in wave 1 (n=4,048). We found significant differences in selection error across modes on all but the 'urbanization' characteristic. However, the small size of *V* (<.10 in all cases) suggests that the category-specific differences in selection error were small across modes.

Table 2.2: Absolute and relative selection error on eight socio-geographical indicators from the national register (Cramér's V)

	Absolute Selection Error (V)				Relative Selection Error (V)
	F2F	Tel.	Mail	Web	All modes
Gender	.046*	.036	.050*	.050*	0.045*
Age	.092**	.125***	.184***	.128***	0.051**
Income	.039	.087**	.141***	.174***	0.050*
Ethnicity	.167***	.228***	.167***	.094***	0.045**
Marital Status	.090***	.149***	.131***	.111***	0.049**
Household size	.078**	.167***	.165***	.139***	0.040*
Urbanization	.171***	.145***	.057	.037	0.036
Urban Cities	.147***	.113***	.038*	.028	0.043*
Sample Size	2,081	2,062	2,182	2,199	4,048

* p<.05; ** p<.01; *** p<.001 (Likelihood ratio test of independence)
Categories of discrete variables:
Gender: 2 categories (male, female)
Age: 6 categories (15-24, 25-34, 35-44, 45-54, 55-64, 65 or higher)
Income: 7 categories (no income, up to 30k Euros, 30-45k, 45-60k, 60-100k, 100k and above, missing)
Ethnicity: 3 categories (Dutch, Western foreigner, Non-Western foreigner)
Marital Status: 4 categories (married or partnership, single, divorced or widowed, missing)
Household Size: 3 categories (One person, two, three or more)
Urbanization: 5 categories (very strong, strong, moderate, little, none)
Urban Cities: 2 categories (Living in Amsterdam, Rotterdam or Utrecht, Living elsewhere)

This analysis is still uninformative about the absolute size of selection error on variable level. For this purpose *V* is estimated, separately for each mode, from the contingency tables of the response indicator and the socio-demographics (formula (2.1)). We found statistically significant selection error in all modes for most variables (column absolute selection error). The size of effects can be classified as 'small' in all cases, however (*V* <.30). The strongest differences in *V* were found for the income distribution and the two regional indicators. On the income variable, web showed strongest selection error followed by mail, whereas telephone and F2F showed weaker or no selection error. To the contrary, F2F and telephone showed stronger error on the

regional indicators. Compared across variables, the risk for selection error implied by the RRs was not reflected by the order of *V*. 'Income' was the only variable showing the expected order (i.e., F2F showed least and web most selection error). Furthermore, the size of *V* statistics indicated small effect sizes throughout, suggesting that no mode stood out strongly in terms of accumulated selection error. Put differently, each mode seemed to have weaknesses and strengths in terms of socio-demographical reach, but differences were generally small.

Subsequently, we extended this analysis to selection error on target variables from the CVS surveyed during the second wave in F2F. Table 2.3 shows the 22 variables sorted in groups of questions on the 'social quality of the neighbourhood', 'problems in the neighbourhood', 'summary ratings about the neighbourhood' and 'factual questions'. The factual questions about past victimization are aggregated from multiple questions about victimization to diverse forms of crimes. As explained in section 2.3, these statistics are based on pooled multiply imputed data sets. Inference to the population is only valid given units were missing at random on socio-demographics. In addition, we assessed selection error against wave 2 F2F respondents only (i.e., deleting wave 2 unit nonrespondents list-wise). The findings of this analysis were equivalent to the ones shown here.

Relative selection error was small and insignificant for nearly all variables except for 'safety feeling' suggesting that any differences in absolute selection error perhaps were likewise small (Table 2.3). Considering absolute selection error, we found that this was indeed true for many variables. However, in some cases *V* in F2F and telephone appeared somewhat higher than in mail and web, suggesting slightly stronger selection error on these variables (e.g., 'safety feeling', 'unpleasant people on the streets').

Table 2.3: Absolute and relative selection error (Cramér's V) on 22 CVS target variables measured in wave 2 (estimates based on ten multiply imputed data sets)

	Absolute Selection Error (V)[a]				Relative Selection Error (V)[a]
	F2F	Tel.	Mail	Web	All modes
Questions 'social quality neighbourhood'[b]:					
State of roads, walkways, and squares	.052	.035	.043	.058	.022
Good playgrounds for children	.039	.060	.030	.041	.031
Good provisions for younger persons	.029	.024	.033	.048	.039
People know each other well	.089**	.089*	.054	.051	.025
People treat each other well	.110**	.148***	.060	.090*	.050
Nice neighbourhood with solidarity	.096*	.055	.060	.051	.032
Feel at home with people	.118**	.090*	.084*	.064	.031
Have a lot of contact with people	.072	.055	.049	.050	.023
Satisfied with population composition	.113**	.106**	.089**	.045	.033
Questions 'neighbourhood problems'[c]:					
Plastering on walls and/or buildings	.096*	.057	.089*	.050	.032
Harassment by groups of young persons	.129***	.102**	.063	.062	.033
Drunken people on the streets	.088**	.083*	.097*	.060	.045
Unpleasant people on the streets	.125**	.099**	.078	.068	.020
Junk on the streets	.094*	.064	.053	.052	.035
Dog excrements on the streets	.031	.042	.036	.071	.035
Destruction of telephone cells, etc.	.050	.029	.028	.023	.022
Drug problem	.121***	.074	.090	.046	.035
Summary ratings about neighbourhood:					
Safety Feeling: insecure[d]	.130***	.059*	.011	.028	.061*
Quality of life rating[e]	.115**	.121***	.104**	.088	.027
Factual Questions (yes / no):[f]					
Police contact (past 12 mth.)	.061	.053	.056	.033	.029
Victim of crime (past 12 mth.)	.075*	.102**	.073	.059	.034
Victim of violent crime (past 12 mth.)	.079	.105**	.064*	.050	.034
Sample Size	2,081	2,062	2,182	2,199	4,048

* p<.05; ** p<.01; *** p<.001 (Likelihood ratio test of independence)

a. Mean weighted estimates of Cramér's V across ten multiple imputations, significance levels based on adjusted Chi² test of independence acc. to Li *et al.* (1991). All questions were surveyed with separate 'Don't know' answer categories. On most variables tthese categories were very sparse (<1% of cases) and were imputed.

b. Question with five rating scale answer categories from 'Completely disagree' to 'Completely Agree' and 'Don't Know'. To avoid cell sparseness, categories 'completely agree' and 'agree' as well as 'completely disagree' and 'disagree' had to be merged in the analyses.

c. Questions with three rating scale answer categories: 'Happens Frequently', 'Happens sometimes', and 'Happens rarely or never'.

d. Question 'Do you sometimes feel insecure?' with two answer categories: yes, no

e. Question on a summary rating about the neighbourhood on a ten point scale with a 'Don't Know' category. To avoid cell sparseness, recoded to four categories (1 to 6, 7, 8, and 9 to 10).

f. Dichotomous measures about police contact within past 12 months and two summary indices based on multiple questions about victimization.

The large number of *V* estimates for socio-demographics and CVS variables and somewhat differing findings across variables still aggravated generalizing conclusions about the impact of mode on selection error. Therefore, it was instructive to additionally assess absolute selection error across these sets of variables (cf. section 2.3.3). We first visualized the empirical distribution of variable-specific absolute selection error using boxplots of the *V* estimates from Tables 2.2 and 2.3 in Figure 2.1. Each boxplot contains the same estimates as listed in these two tables, where triangles denote means and bars medians. This illustration allowed three additional insights: first, error was somewhat larger for the socio-demographic indicators than for the CVS variables. Second, on socio-demographics, mail and telephone evoked somewhat higher selection error, on average, than F2F and web. However, the large ranges of all boxplots indicated strong variance across variables, as noted above. Third, selection error of CVS variables was slightly stronger in F2F and telephone than in mail and web. Again there was considerable variance across CVS variables.

In considering these results it is important to note again the magnitude of all Cramér's *V* estimates indicating small effect sizes (selection error) in all cases. Thus, no mode stood out greatly in terms of absolute error. Remarkably, however, web evoked, on average, less error than telephone and mail despite its lower RR (28.7%).

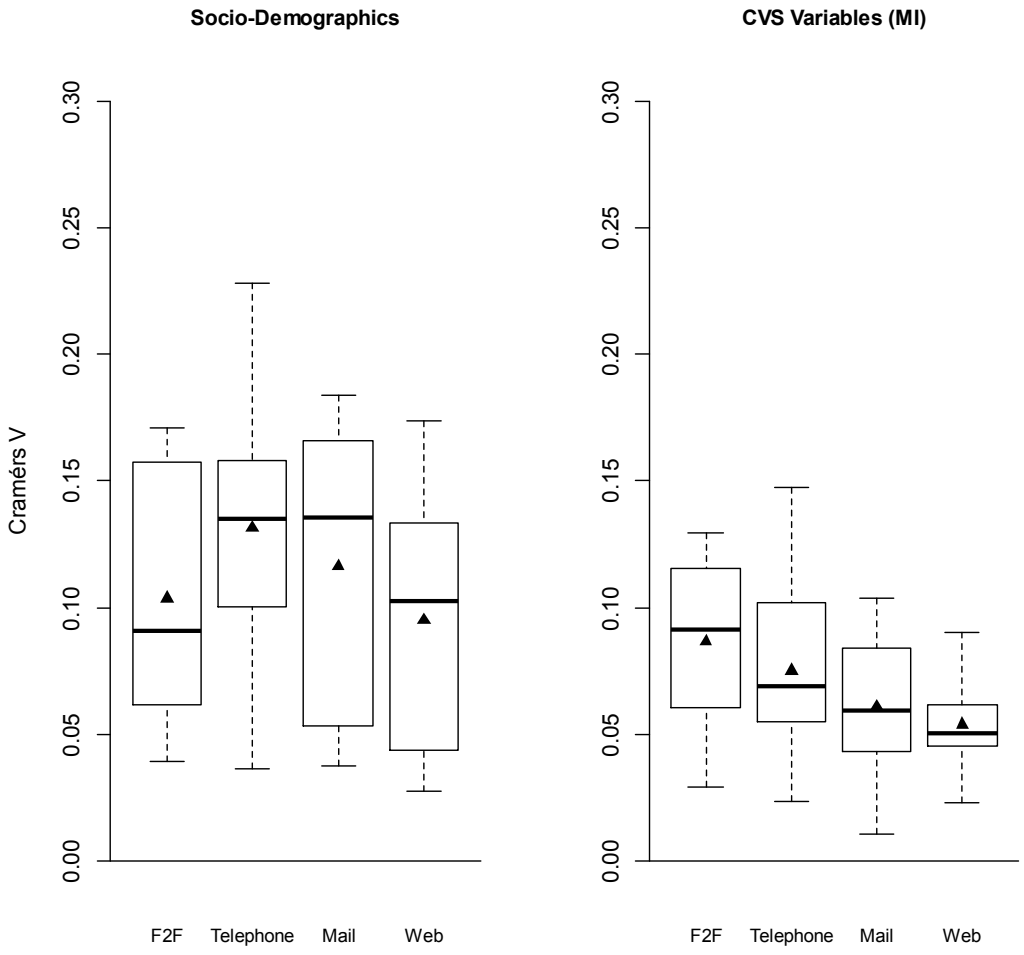


Figure 2.1: Empirical distribution of Cramér's V statistics from Table 2.2 and Table 2.3 (measuring absolute selection error) for the four single-mode designs (triangles indicate means, bars indicate medians)

Finally, we abstracted from the level of variables to a summary score for absolute selection error using the R-indicator (formula (2.4), in short '*R*'). The left part of Table 2.4 shows *R* for the single-mode designs. All estimates were somewhat larger for the CVS variables than for socio-demographics suggesting selection error was smaller, overall, on target variables, which reflected findings from Figure 2.1. Also the order of *R* estimates relates to the findings from Figure 2.1. Web (.798) and F2F (.768) achieved slightly more representative response than telephone (.698) and mail (.719) on socio-demographics. On CVS variables differences were less pronounced, where web (.882) achieved slightly more representative response than telephone (.831) and F2F (.823). However, confidence intervals of these R-indicator estimates overlapped to great extents, suggesting mainly insignificant differences between modes for CVS variables.

Clearly, the order of magnitude of these estimates did not reflect the order of RRs shown again in the last row of Table 2.4.

Table 2.4: Representativeness (R-) indicators for socio-demographics and target variables by single- and mixed-mode designs

	Single Mode Designs				Mixed-Mode Designs		
	F2F	Tel.	Mail	Web	Tel+F2F	Mail+F2F	Web+F2F
Socio-Dem.	.768	.698	.719	.798	.751	.811	.803
	[.728,.809]	[.660,.736]	[.681,.757]	[.759,.836]	[.706,.796]	[.766,.855]	[.757,.850]
CVS Variables	.831	.823	.857	.882	.789	.842	.825
	[.765,.897]	[.761,.885]	[.789,.926]	[.828,.937]	[.731,.846]	[.777,.907]	[.753,.898]
Response rate[a]	64.3	48.2	49.8	28.7	65.2	67.3	59.6

For CVS variables, mean *R* across ten multiply imputed data sets is reported. Variances of *R* were estimated using Rubin's rule for multiply imputed data sets (Rubin, 1987, p. 90).
a. Based on Table 2.1.

2.4.2 Impact of the mixed-mode designs on absolute selection error (RQ2)

The three sequential mixed-mode designs greatly increased RRs by the F2F follow-up survey to approximately the level of a single-mode F2F survey (64.3%; cf. Table 2.4). The second RQ asked, whether this shift in RRs caused a mitigation of selection error. To asses this question, we first considered the change in R-indicators (Table 2.4, right column). On socio-demographics, we noted an upward trend for telephone (.698 to .751) and mail (.719 to .811) indicating a reduction in selection error, but web was unchanged (.798 to .803). However, for the CVS variables we did not find any increased R-indicators. In fact, somewhat reduced *R* for telephone (.823 to .789), mail (.857 to .842), and web (.882 to .825) suggested a slight increase in selection error. It appeared that the mixed-mode follow-up was capable of reducing error in some, but not all cases.

Comparing the mixed-mode designs with the single-mode F2F survey offered an explanation for these findings. The *R* of F2F was higher than telephone and mail for socio-demographics, but it was not higher than web. For CVS variables F2F *R* was slightly smaller than of mail and web, although on insignificant level. Therefore, it

appeared that *R* of the mixed-mode designs trended towards the single-mode F2F survey. This seemed plausible as also the RRs of the mixed-mode designs were increased to the level of F2F but did not clearly exceed it.

We investigated this idea in more detail using the distribution of variable-specific *V*-indices measuring absolute selection error (equation (2.1)) shown in Figure 2.2 contrasting the single- and mixed-mode designs. Clearly, absolute selection error on socio-demographics of mixed-mode mail and mixed-mode telephone showed a downward shift towards the level of single-mode F2F. Web remained unchanged at the level of F2F in central tendency. However, considering CVS variables we found an upward change of *V* statistics in web, which increased to the level of single-mode F2F. These trends were also visible for telephone and mail. It should be noted again that the magnitude of selection error on CVS variables was very small and insignificant in many cases, so that these effects have to be considered as likewise small.

These effects suggest that the absolute selection error of the three single-mode samples was turned more similar to the single-mode F2F sample by the F2F follow-up. However, error was not mitigated on all variables (RQ2). Rather, the F2F follow-up also increased absolute selection error on some of the variables (e.g., CVS variables for web). In Appendix A we illustrate the details of this process. If telephone, mail or web exhibited smaller error than F2F, selection error was increased by the mixed-mode extensions. Furthermore, the mixed-mode follow-up was only capable of decreasing error, if F2F showed smaller error than telephone, mail or web. These effects were found for both types of variables and all mixed-mode designs. Even for the web condition, for which no change in *R* (Table 2.4) and central tendency of *V* (Figure 2.2) was found on socio-demographics, error was changed towards F2F. However, it was both mitigated on some variables and increased on others. Therefore, we could not observe an overall change on systematic level though it became apparent when considering all variable-specific changes.

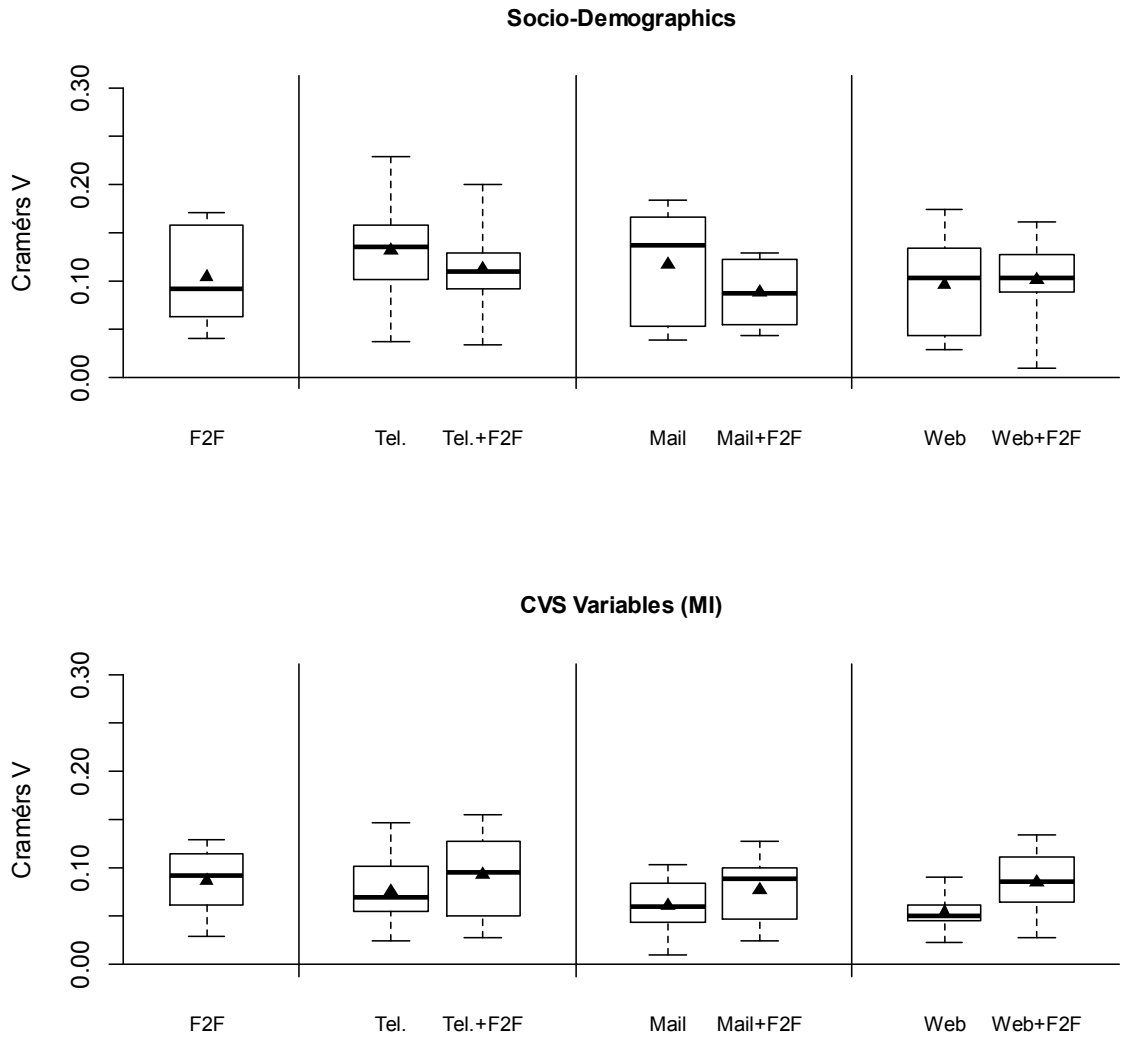


Figure 2.2: Comparison Cramér's V statistics (measuring absolute selection error) between the single- and mixed-mode designs (triangles indicate means, bars indicate medians)

2.5 Discussion

Response and coverage rates are often assumed to indicate the risk for selection error of a survey design and as such they have become established as central indicators of survey quality. Strong enhancements in RRs, furthermore, are often referred to as an advantage of mixed-mode surveys promising to keep selection error small besides reducing costs. For this reason, mixed-mode surveys have gained popularity, especially in face of cost pressure in recent years. The present study assessed the legitimacy of this common practice empirically.

In conclusion, RRs did not indicate the extent of selection error of the single-mode designs (RQ1). In fact, differences between modes were far less pronounced than commonly assumed. There was generally some variance in selection error across variables suggesting that error was rather a function of variables than a function of mode, though absolute selection error was classified as small based on Cramér's V (and for most CVS variables it was practically negligible). The small systematic differences that we still found did not reflect the order of mode-specific RRs (e.g., F2F did not show least and web most error). On socio-demographics, F2F and, surprisingly, web evoked least systematic differences in selection error (greater representativeness as measured by the R-indicator), whereas error on telephone and mail was slightly stronger. On CVS variables, selection error was generally small and often insignificant and mode differences even less pronounced. We still found the tendency for F2F and telephone to evoke somewhat larger selection error than mail and web. The fact that web, in face of its low RR (28.7%), performed slightly better than the other modes on both types of variables is remarkable. We point to the fact that also web respondents were recruited from a person sampling frame and invited by letter to the survey, which probably contributed to the good performance of the mode. More generally, these findings reflect the empirical results of Groves (Groves & Peytcheva, 2008; Groves, 2006) regarding the weak relation of RRs and nonresponse bias.

If these results can be generalized beyond the present survey design to other target variables and populations, selection error may turn out as a less important factor in survey mode choice than commonly assumed and RRs should gain less prominence in arguing for or against particular modes. Furthermore, other quality and cost criterions would need to gain more weight in design decisions. Considering the high costs of F2F or telephone surveys, opting for cheaper modes, in particular web, would be a natural choice. However, further design considerations would need to be balanced against potential cost savings, for example, the extent of mode-specific measurement error and the length and complexity of questionnaires. It appears useful to select modes producing less measurement error, but cheaper modes, such as web, might increase measurement error for some survey topics. Assessing mode differences in measurement error is an important area of further research for this reason.

The answer to RQ2 (i.e., whether RR increases indicated a reduction in selection error) requires more differentiation. The F2F follow-up survey to telephone, mail, and web nonrespondents turned out capable of reducing selection error of the single-mode designs, when the error in the single-mode designs was larger than in F2F. However, when error in the single-mode design was smaller than in F2F, error was enlarged. By this process the small systematic differences between the single-mode designs were reduced in the mixed-mode designs. In this sense, the mixed-mode designs were successful at stabilizing selection error at the level of F2F, which is one goal of mixed-mode surveys. However, on global level a reduction in selection error beyond F2F was not possible and, if F2F showed stronger error, the mixed-mode follow-up could even enlarge this error. These findings imply that the mixed-mode samples could enhance the single-mode samples to reach a certain population domain accessible to F2F surveys, but could not reach different respondents in order to reduce error below that of single-mode F2F. This conclusion is supported by the fact that the mixed-mode RRs could be increased to, but not clearly above the level of single-mode F2F. These observations where most pronounced for the socio-demographic variables, whereas for the CVS variables selection error differences were already very small and these effects were only found in tendency.

Centrally, these results question the current practice of using sequential mixed-mode surveys as a tool for controlling selection error based on the conjecture of increased RRs. Therefore, practitioners need to be sure about the beneficial effects of mixed-mode designs on selection error for particular target variables. When F2F is a priori known to evoke less selection error than a different single-mode design, our results suggest that a mixed-mode follow-up in F2F perhaps can decrease this error, at least as far as socio-demographics are concerned. Otherwise, however, there remains the risk that mixed-mode can even enlarge selection error despite increases in RRs. Furthermore, mixing modes requires additional expenses and presents the possibility of increased measurement error. These facts and the rather small differences in selection error observed in the present study raise doubt, whether the added value outweighs these additional problems of mixed-mode designs.

Having said this, the crucial question is to what extent the present findings can be generalized. Our study represents a single case, which however improved considerably

over prior research (cf. section 2.1.3). Data were based on a high quality probability sample so that generalizations to the Dutch population are admissible. Furthermore, the set of socio-demographic and CVS benchmark variables was exceptionally rich and of high quality. However, conclusions about target variables are based on a MAR assumption used to adjust nonresponse in wave 2. Furthermore, assessing different survey topics might alter conclusions about target variables. This threat appears less likely for socio-demographics, however, which represent a more general class of variables that is, in principle, observable in any survey.

Results may be also affected, if this study is repeated in populations with different survey climates and different coverage by internet or telephone access. In the Netherlands, internet access has reached approximately 90% of households in 2011 (Eurostat, 2012). The good performance of the web mode might therefore be a particularity of The Netherlands. This statement also applies to countries with different landline telephone coverage (71.5% in the present study). Furthermore, mobile phones were not included in the present study and doing so might change results. In all of these respects a reproduction of findings for different survey topics and populations is strongly desirable.

3 Evaluating Mode Effects on Random and Systematic Measurement Error [1]

Analysts of data generated by different modes of data collection, like face-to-face, telephone, paper and pencil ('mail'), or online ('web'), need to be sure that answers to the same questions asked under different modes are equivalent. This concern has gained increased prominence in the context of cross-sectional or longitudinal mixed-mode surveys, in which two or more survey modes are combined concurrently or sequentially to maximize response rates and to optimize on costs (De Leeuw, 2005; Dillman, Smyth, et al., 2009; Dillman, Phelps, et al., 2009). There are in fact strong theoretical arguments that social and cognitive factors impact the answering processes to very different extents (Bowling, 2005; De Leeuw, 2008; Dillman, Smyth, et al., 2009, pp. 311–329; Tourangeau et al., 2000). A sizable body of empirical studies has assessed measurement effects of survey modes, also referred to as 'mode effects', from mainly two perspectives. The 'sampling statistics approach' seeks to assess measurement effects by testing differences in marginal means and, more rarely, variances of separate questions (Link & Mokdad, 2005; Schonlau et al., 2004). The 'answering behavior approach' considers differences in indicators of answering behavior, such as acquiescent, extreme, nondifferentiated, socially desirable or 'don't know' answering (Chang & Krosnick, 2009; Christian, Dillman, & Smyth, 2008; De Leeuw, 1992; Fricker et al., 2005; Greene et al., 2008; Holbrook, Green, & Krosnick, 2003).

[1] This chapter has been published as: Klausch, T., Hox, J. J., & Schouten, B. (2013). Measurement Effects of Survey Mode on the Equivalence of Attitudinal Rating Scale Questions. *Sociological Methods & Research, 42*(3), 227–263. doi:10.1177/0049124113500480

In this paper we follow a third approach to studying measurement effects of modes on attitudinal questions that defines equivalence as independence of answers to a question from a survey mode conditional on latent true scores (Mellenbergh, 1989; Meredith, 1993; Millsap, 2011). This definition implies that two persons with the same true state on the concept of interest give a particular answer with the same probability when asked under different modes. A model-based approach using ordinal multiple-group confirmatory factor analysis (MCFA) is applied to describe how modes impact this probability differentially, referred to as measurement effects (Alwin, 2007; Bollen, 1989; Jöreskog, 1971a; Kankaraš, Vermunt, & Moors, 2011; Millsap, 2011; Skrondal & Rabe-Hesketh, 2004).

A limited number of prior studies have used continuous MCFA approaches to study mode equivalence (Buchanan, Johnson, & Goldberg, 2005; Cole, Bedeian, & Feild, 2006; De Beuckelaer & Lievens, 2009; De Leeuw, Mellenbergh, & Hox, 1996; Deutskens, de Ruyter, & Wetzels, 2006). One potential problem of this literature is that the ordinal measurement level of the attitudinal questions generally is not taken into account. Continuous MCFA models then are error prone in the detection of measurement effects, because ordinal data violate distributional assumptions (Alwin, 2007; Kankaraš et al., 2011; Millsap, 2011, pp. 26–37). More sensible alternative model choices are latent trait models using appropriate link functions (e.g. probit as in ordinal MCFA) or categorical latent traits (Kankaraš et al., 2011; Kim & Yoon, 2011; Meade & Lautenschlager, 2004). Ordinal MCFA, applied in the present study, can be considered a generalization of polytomous item response theory (IRT) models allowing estimation of error variance, which is not possible in IRT (Kamata & Bauer, 2008; Millsap & Yun-Tein, 2004; Millsap, 2011; Muthén & Asparouhov, 2002; Muthén & Muthén, 2010; Muthén, 1984).

Another limitation of prior work is neglecting to take selection effects of modes into account. Practical implementations of mode experiments have shown that sample compositions often are not homogenous across conditions, even if randomization is used (Dillman, Phelps, et al., 2009). This causes a counterfactual situation (Morgan & Winship, 2007), in which it is unknown, whether an observed measurement effect is caused by the mode or the selection process (Jäckle et al., 2010; Vannieuwenhuyze &

Loosveldt, 2013). We apply a propensity score adjustment method in the estimation of the ordinal MCFA model to control for the selection problem.

MCFA allows estimating three types of measurement effects. First, modes may cause differences in the *scale* of a given item that is sensitive to mode by altering the relationship of the true score and the observed answer, i.e. in expectation the same respondent would not give the same answer when asked under different modes (Millsap, 2011, pp. 5–7; Vandenberg & Lance, 2000). Second, modes may change the extent of *random measurement error* of an item sensitive to mode implying differential reliability (precision) and thus attenuated relationship estimates, though the answer probabilities are unbiased (Biemer & Stokes, 1991; Fuller, 1987). Third, a previously rather neglected advantage of MCFA is estimation of relative differences in the extent of *systematic error*[2] across sets of questions, also called nonrandom errors, correlated errors and method variance (Alwin, 2007, pp. 41–42; Andrews, 1984; Blalock, 1970; Davis, 1997; Gerbing & Anderson, 1984; Green & Citrin, 1994; Saris & Andrews, 1991). These are person- and method-level sources of *systematic bias* and *systematic variance* affecting all indicators equivalently.

The MCFA approach thus allows additional insights over the sampling statistics and the answering behavior approaches. Marginal analyses of means and variances can neither distinguish item-specific scale bias from systematic bias, nor can item-specific random error be differentiated from true score variance and nonrandom error variance. The answering behavior approach, furthermore, has described many types of behaviors, but cannot estimate their statistical effects. Answering behaviors are likely causes of systematic error, however. For example, nondifferentiation (Krosnick, 1991) or acquiescence (Billiet & McClendon, 2000) are causes of systematic bias and variance not accounted for by the true score. MCFA thus establishes a model-based link between the answering behavior and the sampling statistics approaches.

---

[2] In the context of this chapter the term 'systematic error' concerns a person-level source of bias an variance that is attributed to all questions in a scale regardless of content (bias or variance). It should be noted that 'systematic measurement error' in the total survey error framework introduced in chapter 1 concerns the bias of a single variable against the true score. In the psychometric models (CFA models) applied in the present chapter this error is represented by the item scale parameters (loadings, thresholds or intercepts) and relative differences in item bias can be tested by comparing scale parameters across mode-specific measurement models, as described in section 3.4.1.

In practice, it is very relevant to know whether measurement effects are item-specific or systematic phenomena. Presence of different extents of systematic error signifies different relationships of true score and observed answers of all items and hence systematically incomparable modes. Such a difference would indicate, for example, that wording and topic of a question are less important in influencing measurement effects. However, if measurement effects only concerned single items these could be taken into account in design, e.g. by changing wording, or in analysis, e.g. by allowing for partial non-equivalence or by indicator omission from analyses that require pooling of data across modes.

Our data stem from a large-scale mode experiment based on a probability sample from the general population of The Netherlands implying high external validity (gross n=8,800). Earlier MCFA studies mainly considered special interest groups[3]. Furthermore, the data allow a comparison of the four major survey modes (F2F, telephone, paper, and web). In prior literature pairwise comparisons of web and paper modes prevailed.

We proceed by our expectations about measurement effects (section 3.1), followed by a description of our experimental data (section 3.2). Section 3.3 introduces the technical background of the methodology. Section 3.4 presents results on three scales. We discuss and conclude by section 3.5.

3.1 Expectations about Measurement Effects of Modes

Historically, a prominent cause of measurement effects of modes was rooted in different traditions of questionnaire design. To eliminate this alternative explanation, researchers should apply 'unified designs' of questions across modes suggesting, for example, equal wording of questions and labeling of answer scales (Dillman, Smyth, et al., 2009, pp. 321–329). Any remaining effects of factors that cannot be equalized are usually

[3] With exception of De Leeuw, Mellenbergh, & Hox (1996) who used a telephone book sampling frame in The Netherlands. Our study however used an address registry sampling frame, which can be considered to cover the general population more fully, because of absence of Telephone coverage bias.

attributed to mode (Groves et al., 2010, pp. 160–162), of which there are two major ones.

First, the social situation during the answer process naturally differs caused by the presence of an interviewer, in telephone or face-to-face modes, or its absence, in self-administered paper and pencil or web modes (Bowling, 2005; De Leeuw, 2008; Dillman, Smyth, et al., 2009, pp. 311–314; Tourangeau et al., 2000, pp. 289–312). A well-known consequence of this factor is socially desirable answering in interviewer administered surveys. But interviewers can also provide motivation in the answer process and throughout the interview, can probe answers, clarify and reassure that respondents focus on the interview. These aspects can enhance attention and depth of cognitive processing. Interviewers, however, are in control of the pace of the interview and the order of questions (Bowling, 2005; Holbrook et al., 2003). This may not give respondents sufficient time to consider answers thoroughly. Especially from telephone surveys it is known that respondents may feel pressured to answer questions, because pauses are perceived as undesirable (De Leeuw, 2008). Conversely, self-administered surveys allow a self-chosen pace and order in an anonymous situation, but lack interactive advantages of motivation and clarification.

The second major difference lies in primarily aural or primarily visual communication of questions and answers (De Leeuw, 2008; Dillman, Smyth, et al., 2009, pp. 314–320; Tourangeau et al., 2000, pp. 289–312). In aural-based modes, question and answer categories need to be fully memorized, whereas in visual modes respondents can re-read question elements multiple times. These tasks pose very different cognitive demands and burden (Bowling, 2005; Fricker et al., 2005; Greene et al., 2008; Heerwegh & Loosveldt, 2008). This difference in cognitive stimulus is likely to impact the full answer process.

On the surface, interviewer modes, like face-to-face (‘F2F’ in the following) or telephone, have very similar measurement properties, because interviewers are present, and they both rely on aural information transmission (supposing no visual elements are used in F2F, as in our study). Self-administered modes, like web and paper and pencil (‘paper’ in the following), are similar due to self-administration and reliance on visual information transmission. Consequently, the measurement processes of interviewer and

self-administered modes differ strongly. It can therefore be generally expected to find small or no measurement effects when comparing telephone with F2F (*Hypothesis H1a*) and web with paper modes (*Hypothesis H1b*), respectively. Moreover, measurement effects should primarily be present between the two interviewer and self-administered modes (*Hypothesis H2*). We posit these hypotheses for questions' scales as well as random measurement errors.

In empirical studies using MCFA modeling H1a has been supported in comparisons of web and paper-based surveys, which found these modes to be fully equivalent (Buchanan et al., 2005; Cole et al., 2006; De Beuckelaer & Lievens, 2009; Deutskens et al., 2006). Yet, since the populations of these studies were rather specific (employees in national or international businesses), external validity is not fully assured. De Leeuw and colleagues (1996) additionally assessed equivalence with respect to interviewer modes based on a national RDD survey with telephone, F2F and paper modes. They find non-equivalence across all modes, in particular between paper and the two interviewer modes (consistent with hypothesis 2), but also the telephone and F2F modes were not fully equivalent. The study by De Leeuw et al. thus points to a potential challenge to H1b. For an indication about random error differences across modes, we refer to a meta-analysis of multi-trait-multi-method (MTMM) literature by Saris and Gallhofer (2007). The authors report that reliability of measurement differs between interviewer and self-administered surveys, which is consistent with H2. In particular, reliability was lower in interviewer administered surveys (cf. Braunsberger, Wybenga, & Gates (2007) for similar results from a web-telephone comparison).

Measurement properties of modes, e.g. differential demands of the social situation, motivation and cognition, are known to impact the occurrence of answering behaviors as well (Heerwegh & Loosveldt, 2008; Holbrook et al., 2003). Studies describing answering behavior found differences primarily between interviewer and self-administered modes, which conforms to hypothesis 1 and 2. Four consistent findings are particularly worth noting. First, if the construct of interest is sensitive, social desirable answering probably affects all indicators of a scale more strongly in the interviewer modes. Second, aural modes have been reported to yield more extreme or more extreme positive responses independent of question content (Christian et al., 2008; De Leeuw, 1992, 2008; Dillman, Smyth, et al., 2009, pp. 316–320; Dillman, Phelps, et al., 2009).

Third, acquiescent answering behavior was found more often in telephone than in web surveys (Greene, Speizer, & Wiitala, 2008; Holbrook, Green, & Krosnick, 2003; however cf. Heerwegh & Loosveldt, 2011). Fourth, nondifferentiation and 'straight lining' answering behaviors have been found to differ between interviewer modes and web surveys (Chang & Krosnick, 2009; Fricker et al., 2005; Greene et al., 2008; Heerwegh & Loosveldt, 2008; Holbrook et al., 2003). These empirical results suggest that the extent of systematic error might also be structured as posed by hypotheses 1 and 2, i.e. primarily be present between interviewer and self-administered modes (*Hypothesis H3*) as sources of a difference in the extent of systematic bias (*H3a*) and variance (*H3b*).

On the empirical side, systematic variance differences between modes are found in the meta-analysis of Saris and Gallhofer (2007). Interviewer administered surveys appear to create higher systematic variance than self-administered surveys giving some empirical support to H3b. Heerwegh and Loosveldt (2011) report on a systematic bias between a telephone and a paper survey, which is interpreted as social desirability effect, consistent with H3a.

3.2 The Dutch Crime Victimization Survey Mode Experiment

Data are available from a mode experiment in The Netherlands conducted from April to June 2011 by Statistics Netherland. The topic and large parts of the questionnaire were adopted from the national Crime Victimization Survey (CVS), an existing cross-sectional survey conducted on a yearly basis by Statistics Netherland. The experiment was administered independently from the regular CVS at a different time and with a different sample. A simple random sample of 8,800 persons was drawn from the national address register and each person was randomly assigned to one out of four modes. 8,524 persons were eligible: 2,081 in F2F, 2,062 in telephone, 2,182 in Paper, and 2,199 in web. All persons received mailed pre-notifications and multiple reminders, where self-administered modes additionally contained either a link to a web survey or a paper questionnaire with a return envelope. In the interviewer modes, contact was attempted by telephone or in person. 4,048 respondents participated. AAPOR Response

Rates 1 were: F2F 64.3% (1,338), telephone[4] 67.4% (993), paper 49.8% (1,086), web: 28.7% (631).

Statistical analyses were conducted on three unidimensional scales. Questions, item wording, and answer categories are shown in Table 3.1. Two scales were based on indicators that are regularly included in the CVS ('Neighborhood Traffic Pressure', NTP, and 'Police Visibility', PV, both 4 indicators). These were explored and cross-validated on a different data set, the web version of the regular CVS from 2010. The third scale ('Duty to obey the police', DTO, 3 indicators) was validated in the pretest of the fifth round of the European Social Survey (in F2F). It is normally not included in the CVS.

Ordinal rating scales contained either three (NTP) or five answer categories (PV, DTO). In the interviewer modes answer categories including 'don't know' options were read out once at the outset of a set of questions and repeated upon request. No show cards were used in the F2F mode. In both self-administered modes, indicators were presented in grids with labeled scales. A well-known problem in unified mode designs is the presentation of 'Don't know' (DK) categories (Dillman, Smyth, et al., 2009, p. 327). If offered visually with each question, DK categories are more prominent in web or paper questionnaires than in telephone or F2F. This can affect the 'visual' scale midpoint (Tourangeau, Couper, & Conrad, 2004). Also typically, this leads to more frequent use of DK categories in self-administered modes (De Leeuw, Hox, & Scherpenzeel, 2011). But omitting DK fully might provoke false or random answers in web and paper. Differential presence of DK answers thus is an alternative explanation to the measurement effects we seek to identify. To control for the impact of DK, the treatment of DK categories was varied across the three scales. In the standard CVS, DK categories are explicitly offered in the two self-administered modes. PV and NTP thus had explicit DK in web and paper (Table 3.1). NTP was selected, because it was known from earlier rounds of CVS that DK could be expected at low rates in all modes, despite the visual presence of a DK option in web and paper. In the PV scale higher DK response in web and paper was expected, which is the more common case when

[4] 588 persons without telephone access in the Telephone condition were excluded before computing AAPOR RR1. Non-covered units are not taken into account in the calculation of RR1. Relative to all eligible units the response rate in Telephone was 48.2%.

offering explicit DK. As a contrast, no DK option was offered for the third scale in web and paper (DTO), which is normally not done in the CVS. In paper, only omission could thus lead to item nonresponse, which consequently was low. The online routine required all indicators to be answered. Therefore, DK was fully absent in web on this scale.

Table 3.1: Overview on indicators and scales with 'Don't know' (+ indicator refusal) rates (in %)

	F2F	Telephone	Paper	Web
Neighborhood Traffic Pressure (NTP), early position[1]				
1. Aggressive behavior in traffic	0.4	0.6	8.3	5.9
2. Traffic noise nuisance	0.0	0.1	5.3	1.9
3. Speeding in traffic	0.4	0.5	4.6	1.7
4. Parking problems	0.3	0.2	5.3	2.7
Police Visibility (PV), middle position[2]				
1. The police offer protection to people in this neighborhood.	4.6	3.1	23.2	16.8
2. The police have contact with people from this neighborhood.	9.0	6.8	27.2	26.0
3. The police react to problems in this neighborhood.	10.2	7.3	30.5	24.7
4. The police do their best in this neighborhood.	9.3	5.8	30.5	24.9
Duty to obey the police (DTO), late position[3]				
1. Support the decisions of the police, also if I disagree.	1.8	3.6	2.7	0.0
2. Do what the police say, also if I disagree.	1.4	3.6	3.0	0.0
3. Do what the police say, also if I am treated unpleasantly.	1.7	4.4	2.9	0.0
Sample size (n):	1338	993	1086	631

Scale labels:
1) Question: How often does the following happen in your neighborhood? Answer Categories: Happens almost never or never (1), happens sometimes (2), Happens frequently (3), Don't Know
2) Completely disagree (1), disagree (2), neutral (3), agree (4), completely agree (5), Don't Know
3) Fully not my duty (1), 2, 3, 4, Fully my duty (5)

3.3 Statistical Methodology and Assumptions

This section describes the statistical methodology applied in the analysis of the three scales introduced above. First, we show how the ordinal multiple group confirmatory factor analysis (MCFA) model can be applied for modeling measurement of true scores by categorical indicators under different survey modes (section 3.3.1). Second, the role

of systematic measurement error across sets of questions is discussed in the MCFA model (section 3.3.2). Thirdly, the role and adjustment of selection effects in mode experiments is addressed (section 3.3.3).

3.3.1 The ordinal MCFA model

Ordinal MCFA assumes that the observed random vector of ordinal response variables $\boldsymbol{Y}$ (indicators or items) has a latent response variable vector $\boldsymbol{Y}^*$ linked to it (Millsap, 2011, pp. 122–124). We consider the case when the categorical factor model is estimated in multiple groups (or 'populations'), which in the present analysis are represented by a set of mode-specific samples (groups) *M*. These groups only differ by the measurement instrument (mode) used to administer the same questions. It is assumed that $\boldsymbol{Y}^*$ follow a multivariate normal distribution in each mode:

$$P(Y^*|M) \sim MVN\big(E(Y^*|M), Var(Y^*|M)\big). \tag{3.1}$$

The latent response variables are linked to the observed ordered categorical indicators $\boldsymbol{Y}$ by means of indicator-specific threshold parameter. Suppose there are c=1,…,C categories on each Y, then C-1 threshold parameters are defined as v_c^m. The bivariate normal density between two observed indicators in mode *M* is:

$$P\big(Y_j = c, Y_{j'} = c' \big| M\big) = \int_{v_c^m}^{v_{c+1}^m} \int_{v_{c'}^m}^{v_{c'+1}^m} N\Big(Y_j^*, Y_{j'}^* \Big| M\Big)\, \mathrm{dY_j^*}\ \mathrm{dY_{j'}^*}. \tag{3.2}$$

A key parameter in each bivariate normal density is the correlation, which is estimated for each pair of indicators as polychoric correlation given threshold estimates in each group. The thresholds and mode-specific polychoric correlation matrices form the basis for further mean and covariance structure analysis, as discussed for the case of continuous data in the congeneric factor model (Jöreskog, 1971a, 1971b; Lord & Norvick, 1968). Contrary to continuous factor analysis, the ordinal MCFA model specifies a congeneric factor model for the conditional expectations and variances of the *latent* response variables:

$$E(Y^*|T, M) = \tau^M + \Lambda^M T \tag{3.3}$$

$$Cov(Y^*|T, M) = \Theta^M \tag{3.4}$$

with $\boldsymbol{T}$ an $R\times 1$ vector of common factor scores (true scores) for latent variables $r=1,...,R$, $\boldsymbol{Y}^*$ a $J\times 1$ vector of J latent response variables (indicators), Λ^M a $J\times R$ matrix of $\lambda_{j|M}$ parameters called factor loadings, $\boldsymbol{\tau^M}$ a $J\times 1$ vector of intercept parameters, and $\boldsymbol{\Theta^M}$ a diagonal $J\times J$ matrix of item-specific error variance parameters $\theta_{j|M}$. M identifies separate parameter sets over modes. The unconditional expectations and covariances of M are:

$$E(Y^*|M) = \tau^M + \Lambda^M \kappa \tag{3.5}$$

$$Cov(Y^*|M) = \Lambda^M \Phi \Lambda^{M'} + \Theta^M. \tag{3.6}$$

$\boldsymbol{\kappa}$ is the vector of population means and $\boldsymbol{\Phi}$ population variance-covariance matrix of T. Note that, since we seek to measure the same true score distribution in all modes, population means and variances do not depend on mode.

Thresholds, intercepts and loadings set the scale of each question Y in a given mode. Random errors are explicitly included in the model and can be tested for equivalence. This distinguishes the ordinal MCFA model from other latent variable models for categorical outcomes which are offered, for example, by item response

theory (IRT)[5]. In absence of measurement effects on scale or random error, the parameters are equivalent, which is testable on empirical data. Random error variance is essential for the estimation of indicator reliability (Alwin, 2007), which is expressed as:

$$\rho_{j|M} = \frac{\lambda_{j|M}^2 \Phi}{\lambda_{j|M}^2 \Phi + \theta_{j|M}} \tag{3.7}$$

with $\theta_{j|M}$ error variance of indicator j. Obviously, random error equivalence is only identical to reliability equivalence in the presence of loading equivalence.

Thresholds and intercepts are not simultaneously identifiable, which is why estimation has to focus on one of the two (Millsap, 2011, pp. 128–131). We focus on thresholds in the following by constraining intercepts to zero. All models are estimated using mean and variance adjusted weighted least squares, WLSMV (Asparouhov, 2005; Millsap, 2011, pp. 131–136; Muthén, du Toit, & Spisic, 1997; Muthén, 1984, p. 19). This procedure estimates thresholds, loadings, factor means, factor variances, and error variances simultaneously in all mode-specific response groups by a stepwise procedure. First, thresholds estimates under a multivariate normal distribution are found. Second, polychoric correlations between each pair of indicators are estimated given threshold estimates from the first step. Finally, generalized least squares (GLS) estimation minimizes the discrepancy between the threshold and polychoric correlations estimated

[5] In IRT, the ordinal indicators are modelled without explicit error term, by C-1 indicator difficulty parameters b and an indicator-specific discrimination parameter a: $P(Y \geq c|T, M) = F(a(T - b_c))$ where $F(\,)$ is given by the logistic link function. IRT parameters now relate to ordinal MCFA as follows. Suppose a unidimensional latent variable T, then (Kamata & Bauer, 2008; Muthén & Asparouhov, 2002, p. 5):

$$b_{cj}^M = \frac{v_{cj}^M - \lambda_{j|M} E(T|M)}{\lambda_{j|M}\sqrt{Var(T|M)}} = \frac{v_{cj}^M - \lambda_{j|M} E(T)}{\lambda_{j|M}\sqrt{Var(T)}}$$

$$a_j^M = \lambda_{j|M}\sqrt{\frac{Var(T|M)}{\theta_{j|M}}} = \lambda_{j|M}\sqrt{\frac{Var(T)}{\theta_{j|M}}}$$

Error variances $\theta_{j|M}$ scale the IRT discrimination parameter a (though not b), deteriorating it in groups M with higher error variance. Under IRT one would therefore conclude that non-equivalence on a_j^M is present (referred to as differential item functioning, DIF), if either loadings or error variances are non-equivalent. The ordinal MCFA model thus helps differentiating between random error non-equivalence and scaling problems. IRT models essentially assume homogenous error variances across groups, which might be too restrictive in face of well-known group differences on random errors known from continuous MCFA (Muthén & Asparouhov, 2002).

in the first two steps and the model implied estimates. WLSMV estimation represents a special case of GLS, for which the variance-covariance matrix of estimated thresholds and correlation is reduced to the diagonal elements (variances) of the full matrix. This procedure is primarily used to increase stability of estimation. For detailed model identification restrictions see Millsap and Yun-Tein (2004), Millsap (2011, p. 138), and section 3.4. Measurement equivalence of scale and random error can be assessed by constraining parameters equal across modes, as detailed in section 3.4.

3.3.2 Systematic errors across sets of questions in the ordinal MCFA model

In contrast to random error, systematic errors are common sources of variance and bias of a particular method of measurement, like a mode, affecting measurements in equivalent ways. One way to include systematic errors in measurement models is by an additive, *mode-dependent* random variable *S*. Following approaches by Saris and Andrews (1991) and Scherpenzeel and Saris (1997) we assume that these are observed as a compound with *T*, say for person *i*: $T^*_{i|M} = T_i + S_{i|M}$ (cf. Alwin, 2007, pp. 41–42). Latent response indicator *j* can then be described as ($\varepsilon_{ij|M}$ be an error term):

$$y^*_{ij|M} = \lambda_{j|M}\left(T_i + S_{i|M}\right) + \varepsilon_{ij|M}. \tag{3.8}$$

Define mode-specific systematic bias and variance as $E(S|M)$ and $Var(S|M)$. Statistically, the presence of systematic bias and variance changes the response scale altering both the threshold (intercept) and the loading structure by monotonous shifts. Therefore, they represent serious threats to equivalence of measurement. Assume $Cov(T, S) = 0$, then:

$$E\left(Y^*_j\middle|M\right) = \lambda_{j|M}\left(\kappa + E(S|M)\right) \tag{3.9}$$

$$Var\left(Y^*_j\middle|M\right) = \lambda^2_{j|M}\Phi + \lambda^2_{j|M}Var(S|M) + \theta_{j|M}. \tag{3.10}$$

Equivalently, we may introduce a mode-specific systematic intercept with $\lambda_{j|M}E(S|M) = \tau^S_{j|M}$:

$$E\left(Y^*_j\middle|M\right) = \tau^S_{j|M} + \lambda_{j|M}\kappa. \tag{3.11}$$

This reflects that systematic errors can be interpreted as constant shifts of intercepts, weighted by factor loadings. Since in our model intercepts are zero-constrained for identification, see above, a presence of $\tau_{j|M}^{s}$ causes a weighted shift of all thresholds v_{jc}^{m} simultaneously. Similarly, systematic variance inflates the scale metric (loadings) of all indicators by a constant. Let parameter $\gamma_M = Var(S|M)\Phi^{-1}$ scale S to the unit of true score variance, then for all j:

$$\begin{aligned} Var\left(Y_j^* \middle| M\right) &= \lambda_{j|M}^2 \Phi + \lambda_{j|M}^2 \gamma_M \Phi + \theta_{j|M} \\ &= (1 + \gamma_M)\lambda_{j|M}^2 \Phi + \theta_{j|M}. \end{aligned} \tag{3.12}$$

Another interpretation for the term $\lambda_{j|M}^2 \gamma_M \Phi$ is an increase in indicator error covariance due to the presence of a systematic error variance component. Also note that presence of systematic variance necessarily biases estimates of reliability upwards due to overestimation of true score variance, cf. formula (3.7) (Alwin, 2007, p. 42).

In single-group situations systematic errors are not identified, but estimated as true score compound T_M^*. However, relative differences can be estimated in multi-group situations using factorial designs. Introduce κ_M^* and Φ_M^* as expectation and variance of T_M^*. Assuming loading equivalence across modes (testable assumption), it must hold for any two modes (A and B) due to randomization that:

$$\kappa_{M=a}^* = \kappa_{M=b}^* \Leftrightarrow E(S|M = a) = E(S|M = b) \tag{3.13}$$

$$\Phi_{M=a}^* = \Phi_{M=b}^* \Leftrightarrow Var(S|M = a) = Var(S|M = b). \tag{3.14}$$

The mean and variance difference of the estimated compound factor gives an indication for relative differences in systematic errors, because true score distributions should be balanced by means of randomization. This logic forms the basis of a test of equivalence of systematic errors in the presence of loadings equivalence. Noteworthy, we will consider relative difference in systematic error, which is relevant to conclude about equivalence. We cannot conclude about absence of systematic bias or variance using this approach.

The term *S* is used to approximate[6] the systematic effects of answering behaviors on means and variances of sets of response variables (Billiet & Davidov, 2008; Billiet & McClendon, 2000; Heerwegh & Loosveldt, 2011; Morren, Gelissen, & Vermunt, 2011; Welkenhuysen-Gybels, Billiet, & Cambré, 2003). For example, if persons, depending on mode, vary in their propensity to agree to sets of indicators (acquiescence), *S* has variance and non-zero means in the direction of agreeing, introducing systematic bias and variance. A similar argument can be made for social desirable response behavior. If persons in the population vary in their tendency to provide desirable answers across all indicators in a given model, this introduces systematic variance and a bias in the direction of desirable responses. Moreover, if persons provide extreme answers on all indicators, loadings are scaled upward by a constant, because any true score leads to higher (or lower) responses. As shown above, the presence of *S* with a variance is equivalent to a shift in loadings. Also, behaviors like nondifferentiation and straight lining have been discussed to cause 'correlated errors' of indicators (Gerbing & Anderson, 1984; Green & Citrin, 1994). As mentioned above correlated errors in MCFA models are statistically equivalent to the presence of a systematic error term. From this illustration it is apparent that the mode-dependent term *S* denotes a 'net effect' of the many reasons for systematic error differences across modes, but avoids specifying a particular type of behavior as an error source.

S has also been referred to as invalidity effect of the method, invalidating unbiased measurement of the concept of interest (Mellenbergh, 1999; Saris & Andrews, 1991; Scherpenzeel & Saris, 1997). We conceptualize *S* as compound with *T* in this tradition. Then the effect of *S* is mediated by factor loadings (cf. formulas (3.8)-(3.10)). An alternative conceptualization is to assume a direct effect of *S* on *Y* modeling *S* as a second factor with unit constrained loadings (Billiet & McClendon, 2000; Bollen & Paxton, 1998; Welkenhuysen-Gybels et al., 2003). We will return to this alternative option in the discussion (section 3.5).

[6] It should be noted that all of these answering behaviors might have additionally more complex (non-monotone, non-linear) effects on thresholds or loadings or affecting higher order moments (see e.g. Morren et al. 2011). Modeling these as a source of variance and bias therefore is an approximation.

3.3.3 Selection effects in mode experiments

The above considerations were made for a fully randomized experiment. However, full randomization of persons to modes is seldom possible, especially if samples from the general population are concerned, because modes involve differential sampling frame coverage and evoke differential self-selection (Groves et al., 2010, pp. 162–168). Selection effects are an alternative explanation for measurement effects (Jäckle et al., 2010; Vannieuwenhuyze & Loosveldt, 2013), comparable to counterfactual situations in quasi-experiments (Morgan & Winship, 2007). Two counterfactual situations are possible. First, a selection variable *X* (that is, a variable, which causes selection into mode conditions) might be also related to the true score of interest. Second, there might be measurement non-equivalence across classes of *X*. For these reasons, it is necessary to adjust for selection, e.g. by conditioning on *X*. One way to do so is weighting adjustment by the inverse of propensity scores (Guo & Fraser, 2010; Kaplan, 1999; Morgan & Winship, 2007; Rosenbaum & Rubin, 1983; Rosenbaum, 1987). For more than two modes it is advisable to weight to the population[7]. Define a propensity score model as

$$P(R = 1|X, M) = F(X\beta + MX\gamma), \tag{3.15}$$

where *R* is the response indicator, *MX* indicates all interactions of mode indicators and *X* under useful identification constraints and $F(\cdot)$ is the cumulative normal (probit) or logit link function. From this model, propensity scores $\hat{e}(X, M)$ are estimated as the basis of weights $\widehat{w} = \hat{e}(X, M)^{-1}$. This model calibrates the mode-specific response distributions to the population assuming availability of auxiliary variables *X* on the sample level. Weighting adjustment has been integrated in WLSMV estimation of ordinal MCFA (Asparouhov, 2005).

[7] For two modes (*M=0* or *M=1*) it is also possible to weight to a reference mode ('treatment group') by assigning weights $\widehat{w} = \frac{M}{\hat{e}(x)} + \frac{1-M}{1-\hat{e}(x)}$ (Guo & Fraser, 2010, p. 161; Morgan & Winship, 2007, p. 103; Rosenbaum, 1987).

3.4 Data Analysis and Results

This section discusses the technical details of the data analysis, including equivalence tests of parameters, robustness checks (so-called cross-validation and Oort correction), and the weighting adjustment that was applied. Subsequently, we discuss the empirical results from tests of equivalence of scale, random error, and systematic measurement error on the three scales examined in the present study.

3.4.1 Testing procedure

The testing procedure for mode effects on scales, random errors and systematic errors followed a series of steps, graphically illustrated in Appendix B-1. First, so-called configural equivalence models were fit (model 1), which specify only the factor structure with free parameters under minimal identification constraints[8] – hereafter *'MIC'* (Millsap & Yun-Tein, 2004; Millsap, 2011, p. 138). If configural equivalence held, all loadings and thresholds were constrained simultaneously to test scale equivalence across all indicators ($\Lambda^M = \Lambda, \upsilon^M = \upsilon$, Model 2). This is similar to a constrained backward strategy (Kim & Yoon, 2011; Muthén & Asparouhov, 2002; Stark, Chernyshenko, & Drasgow, 2006; Yoon & Millsap, 2007). An advantage of this strategy is that specifying a particular type of MIC is avoided at first, which is useful, because 'wrong' choices (i.e., constraining an unequal loading or threshold equal for identification) can bias parameter estimation. Deteriorating model fit against the configural equivalence model suggests scale non-equivalence of at least one indicator. If fit deteriorated, the location of misfit was determined on either loadings or thresholds. First, all loadings were freed while thresholds were held fixed (Model 2a). This required unit constraining a particular reference loading. Which indicator to choose, is not trivial, however (French & Finch, 2008; Yoon & Millsap, 2007). We compared model fits of all possible anchor indicators and chose the model that maximized fit.

An improvement in fit of model 2a against the constrained scale equivalence model would suggest that non-equivalence of loadings causes (part of) the misfit. If loading

[8] A set of sufficient MIC for all models estimated were: one unit constrained loading of an anchor indicator, one equality constrained threshold (plus a second threshold for the anchor), and one mean fixed to zero in one reference group that had additionally unit constrained error variances. In the configural model, all particular choices of MIC lead to identical fit.

equivalence holds, but scale equivalence not, non-equivalence is perhaps located on thresholds. This was then tested by freeing thresholds while holding loadings fixed (Model 2b). As MIC, always one threshold per indicator needed to be constrained equal, plus a second threshold for the anchor indicator, where again we chose the MIC which maximized fit.

If non-equivalence was found on the loadings or the thresholds, it was assessed, if the expected structure of measurement effects according to hypotheses H1a/b and H2 held. If there was non-equivalence on thresholds, then this implied testing (Model 2b-1):

$$\upsilon^{F2F} = \upsilon^{Tel} \wedge \upsilon^{Paper} = \upsilon^{Web}$$

against the model which kept all thresholds free (Model 2b). If model fit did not deteriorate, the structure predicted by H1a/b and H2 held. Finally, it was assessed which indicators caused the measurement effect by inspecting parameter estimates of thresholds (or loadings).

Next, random errors were constrained across modes ($\Theta^M = \Theta$, Model 3). Error variances were always tested in the scale equivalence model and, if the scale equivalence model did not hold, additionally in a parsimonious model with good fit, because it is uncertain how sensitive error variance tests are to misspecified baseline models. If full equivalence of random errors was rejected, assessment of H1 and H2 was conducted as explained for non-equal thresholds.

To test systematic bias and variance equivalence we used formulas (3.13) and (3.14) suggesting that due to randomization compound factor means and variances are only equal across modes, if systematic errors are equal. If means κ_M^* were equality constrained, any group difference in systematic bias would have caused a differential shift in thresholds (formula (3.11)). If thresholds were also fixed, however, as in the scale equivalence model, a decrease in model fit indicated the presence of a difference in systematic bias (Model 4a). When the compound factor variance Φ_M^* was equality constrained, systematic variance differences manifested in unconstrained loadings or error covariances by a scaling factor (formula (3.12)). Model fit deteriorated, if this was not possible, when loadings were constrained, as in the scale equivalence model (Model 5; error covariances are fixed at zero in all analyses).

3.4.2 Model fit evaluation, Oort adjustment and cross-validation

WLSMV estimation was conducted by the software Mplus 6.1. 'Don't know' answers or refusals are treated as missing completely at random in WLSMV estimation. Change in fit was assessed by adjusted Chi-square difference tests using the 'difftest' option in Mplus (Asparouhov & Muthén, 2006) and the global fit index RMSEA. A significant Chi-square test denotes a significant change in fit. Fit indices like RMSEA are still under testing for ordinal CFA models (Millsap, 2011, p. 136), but in continuous CFA, RMSEA<.05 is considered a good fit. Furthermore, a simulation by Chen (2007) showed for continuous MCFA that a change larger than .01 in RMSEA indicates meaningful change in fit. We took RMSEA as secondary guideline, but primarily relied on the exact Chi-square test.

Two additional measures were taken to assure robustness of our results. Testing a less constrained model against a constrained baseline model with bad fit has been shown to cause inflated type-one error detection rates of non-equivalence (Kim & Yoon, 2011). Such testing, however, is necessarily done, for example, in tests of loading or threshold non-equivalence when scale equivalence is rejected. The authors showed that the so-called Oort (1998) correction normalizes false positive rates in categorical MCFA with WLSMV. Oort's correction is applied to all affected Chi-square tests[9].

Additionally, all models were cross-validated using half of the sample for testing and the second half of the sample for re-testing. Reported model fit statistics are based on the full sample, but results of difference tests are reported only, if cross-validation suggested equal conclusions in both half splits.

3.4.3 Weighting adjustment

Eight socio-demographic variables were available from the national registers: gender, age (6 categories: 15-24, 25-34, 35-44, 45-54, 55-64, 65 or higher), income (7

[9] Let χ_0^2 be the chi-square fit statistics of the fully constrained baseline model, K be the critical value before adjustment, and df_0 the degrees of the degrees of freedom of the fully constrained baseline model then the adjusted critical value is given by (Kim & Yoon, 2011, p. 225):

$$K_{adjusted} = \left(\frac{\chi_0^2}{K + df_0 - 1}\right) K$$

categories: no income, up to 30k Euros, 30-45k, 45-60k, 60-100k, 100k and above, missing), civil status (4 categories: married or partnership, single, divorced or widowed, missing), nationality (3 categories: Dutch, Western foreigner, Non-Western foreigner), household size (3 categories: 1 person, 2, 3 or more), urbanity (4 categories: strong, moderate, little, none), and living in one of the three biggest national cities (4 categories: Amsterdam, Rotterdam, Utrecht, other). Response propensities were estimated from a probit model including all categorical predictors and their interactions with three mode condition indicators. The reference set were all eligible units including units without telephone access (also in the telephone condition). The maximum normalized weight was 1.846, which is not extreme.

3.4.4 Neighborhood Traffic Pressure (NTP) and Police Visibility (PV) scales

Table 3.2 provides test sequences for the NTP and PV scales. Consider first results for the NTP scale. Very low RMSEA and insignificant model Chi-square indicated very good fit of the configural equivalence model (Model 1). The fit of the full scale equivalence model strongly deteriorated, however, suggesting that scale parameters were non-equivalent (model 2). Freeing loadings across conditions did not result in a significant increase in fit (Model 2a) indicating that the major source of misfit was located on the thresholds. Consequently, we freed all thresholds, while holding loadings equal (Model 2b). Compared to the scale equivalence model, a highly significant Chi-square test and negative RMSEA difference indicated a strongly improved fit. Now we tested H1 and H2 simultaneously by imposing '$v^{F2F} = v^{Tel} \wedge v^{Paper} = v^{Web}$' on the thresholds (model 2b-1). This model did not fit significantly worse than the full non-equivalence model 2b in terms of RMSEA and chi-square value, which suggests that most non-equivalence lies between interviewer and self-administered modes (H2). From these results, we can conclude, consistent with hypotheses 1 and 2, that the non-equivalence of scale was located between interviewer and self-administered modes only, and more particular, on non-equivalent positions of thresholds.

Table 3.2: Equivalence test sequences for the NTP and PV scales

			Neighbourhood Traffic Pressure Scale (NTP)					Police Visibility Scale (PV)				
Model	Equivalence Test	Tested against	RMSEA	RMSEA Diff.	Model Chi²	Adj. Chi² diff. test	Model df (diff.)	RMSEA	RMSEA Diff.	Model Chi²	Adj. Chi² diff. test	Model df (diff.)
1	Configural	-	.014	-	9.6 (n.s.)	-	8	.048	-	25.4[c]	-	8
Scale Equivalence:												
2	Fixed Scale	1	.074	+.060[a]	173.7***	149.8***	26 (18)	.052	+.004	180.2***	156.6***	50 (42)
2a	Free Loadings	2	.085	+.011[a]	141.5***	52.8 (n.s.)[b]	17 (9)	.058	+.006	171.2***	13.0 (n.s.)	41 (9)
2b	Free Thresholds	2	.045	-.029 [a]	50.9***	143.9***	17 (9)	.026	-.026[a]	28.2*	151.8***	17 (33)
2b-1	Web=Paper≠F2F=Tel	2b	.040	-.005	60.8***	9.9 (n.s.)	23 (6)	.033	+.007	79.4***	51.7[c]	39 (22)
Random Error Equivalence:												
3a	Fixed Error Variances	2	.080	+.006	281.0***	121.2**[b]	38 (12)	.059	+.007	270.5***	108.4***	65 (15)
3a-1	Web=Paper≠F2F=Tel	2	.065	-.009	189.8***	5.8 (n.s.)	34 (8)	.052	-/+ 0	206.6***	29.5 (n.s.)[b]	58 (8)
3b	Fixed Error Variances	2b-1	.056	+.016[a]	145.5***	94.6***	35 (12)	.037	+.004	117.4***	42.3*[b]	51 (12)
3b-1	Web=Paper≠F2F=Tel	2b-1	.034	-.006	67.7***	6.5 (n.s.)	31 (8)	.036	+.003	106.2***	31.2 (n.s.)[b]	47 (8)
Systematic Error Equivalence:												
4a	Fixed Factor Means	2	.092	+.018[a]	276.7***	68.5*[b]	29 (3)	.072	+.020[a]	316.8***	60.4**[b]	53 (3)
4a-1	Web=Paper≠F2F=Tel	2	.070	-.004	166.5***	7.0 (n.s.)[b]	28 (2)	.043	-.009	141.7***	3.0 (n.s.)	52 (2)
4b	Fixed Factor Means	2b-1	.047	+.007	86.4***	21.3* [b]	27(4)	.043	+.010	115.3***	19.9**[b]	42 (3)
4b-1	Web=Paper≠F2F=Tel	2b-1	.041	+.001	69.2***	9.4 (n.s.)[b]	26 (3)	.027	-.006	68.5**	3.2 (n.s.)	41 (2)
5	Fixed Factor Variance	3b-1	.040	+.006	89.8***	16.9 (n.s.)[b]	34 (3)	.029	-.007	89.7***	4.6 (n.s.)	50 (3)

NTP Scale: N=4021 (27 cases excluded with Don't know / Refusal on all indicators)
PV Scale: N=3799 (249 cases excluded with Don't know / Refusal on all indicators)
* p<.05, ** p<.01, *** p<.001
'≠' denotes free parameters for two modes, '=' denotes fixed parameters for two modes.
a. Meaningful change of RMSEA criterion (i.e. >.01)
b. Oort adjustment of critical value resulted in a lower significant level or insignificant test. In all other cases adjustment did not change level of significance (Kim & Yoon, 2011; Oort, 1998).
c. Effect / Significance did not hold to cross-validation in both split half samples.

Next, measurement effects on random error were considered suggesting fixing all error variance matrices across modes. Besides the scale equivalence model (2), which had bad fit, error equivalence was also assessed in the threshold non-equivalence model (2b-1), which had improved fit (yielding models 3a and 3b). Results were robust with regard to the type of base model. Both suggested decrease in fit and thus unequal error variances. Subsequently, we imposed the structure implied by H1 and H2 simultaneously, leaving separate error matrices only for the self- and interviewer administered conditions (models 3a-1 and 3b-1). Now model fit did not deteriorate at all, supporting H1 and H2 also for error variances (model parsimony even leads to lower RMSEA).

The findings so far suggest measurement effects on scale and random errors between interviewer and self-administered modes. It was therefore relevant to assess which indicators were affected by measurement effects. Consider the parameter estimates of free thresholds and random errors in model 3b-1 shown in Table 3.3. Surprisingly, threshold estimates for web/paper are consistently lower than F2F/telephone for all indicators (except the thresholds separating categories 'never' and 'sometimes', which are fixed as MIC). Furthermore, consider relative sizes of the item specific error variances. Here we find again a difference on all indicators, where web/paper showed consistently less random error (F2F/telephone fixed at one as MIC).

Table 3.3: Threshold and error variance estimates for the NTP scale (from model 3b-1) with bootstrapped S.E. (10,000 draws)

	Free Threshold (1) (Never/Sometimes)		Fixed Threshold (2) (Sometimes/Frequently)		Random Error Variance	
	F2F/Tel	Paper/Web	F2F/Tel	Paper/Web	F2F/Tel	Paper/Web
Indicator 1	0.157	0.032	1.672	1.672	1	0.515
	(.050)	(.061)	(.078)	(.078)		(.076)
Indicator 2	0.739	0.190	1.475	1.475	1	0.729
	(.049)	(.049)	(.066)	(.066)		(.082)
Indicator 3	-0.724	-0.724	0.679	0.679	1	0.330
	(.058)	(.058)	(.055)	(.055)		(.057)
Indicator 4	0.165	-0.103	0.689	0.689	1	0.703
	(.029)	(.032)	(.032)	(.032)		(.091)

Finally, we tested equivalence of systematic bias and variance. Constraining factor means across modes in the scale equivalence model strongly deteriorated fit providing evidence for different extents of systematic bias (Model 4). Subsequently, we tested, in line with H3a, change of fit when only constraining web to paper and F2F to telephone, respectively (Model 4-1). Doing so did not cause deterioration of fit, suggesting that systematic bias was equal for these modes, but differed between interviewer and self-administered modes (support of H3a). As we found threshold non-equivalence, we additionally tested equivalence of factor means in models with free thresholds (yielding models 4b and 4b-1) leading to the same result.

To assess equivalence of systematic error variance, factor variances were constrained equal based on model 3b-1, which is a parsimonious model with the best fit in terms of RMSEA (Model 5). Model fit slightly deteriorated compared to 3b-1 (RMSEA +.006) and the Chi-square difference test was insignificant, although it was still close to significance (Oort adjusted critical value: 17.2). In the cross-validation one half-split sample was very far from significance, however. This is too little evidence to conclude on the presence of differential systematic variance, rejecting H3b. We also assessed factor variance equivalence in the scale equivalence model (2), leading to the same conclusion (not shown).

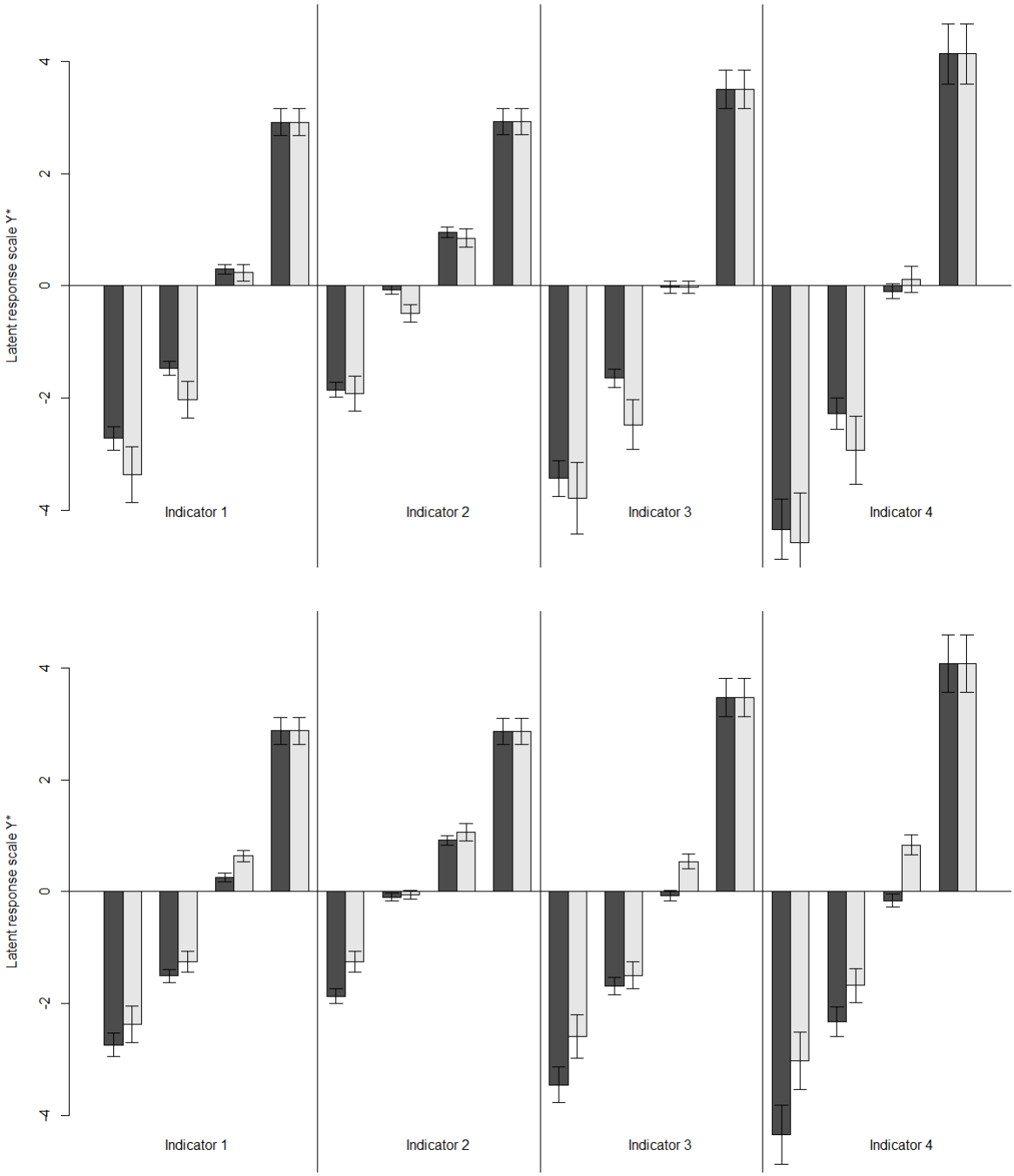


Figure 3.1: Threshold estimates with bootstrapped 95% CI (10,000 draws) for the PV scale (upper part, based on model 3b-1), and with additionally zero-constrained factor means (lower part) illustrating mediated impact of systematic bias (Black: F2F/telephone; Grey: web/paper)

Subsequently, we assessed whether it was possible to reproduce the results of the NTP scale on the police visibility scale (PV). This was possible, without exception, despite the fact that this scale used a different number of answer categories (five instead of

three), had a later position in the questionnaire, and that there were more 'Don't know' answers in the web and the paper conditions. Scale equivalence across all modes was rejected (model 2), while again loading equivalence was not the source of non-equivalence (model 2a). Freeing all thresholds improved model fit strongly (model 2b), where again differences were located between interviewer and self-administered modes only (model 2b-1). Again threshold differences were present on all indicators. The upper part of Figure 3.1 displays four ordinal thresholds for each indicator (based on model 2b-1). For the exact wording of the four indicators we refer again to Table 3.1. In particular the second threshold separating 'disagree' from 'neutral' was always lower in web/paper. Also random error was consistently smaller in web/paper on all except the third indicator, where it was equal (not shown; cf. Table 3.5 discussed below).

There was again a difference in systematic bias, but not in systematic variance (Models 4 and 5). We illustrate the impact of the difference in systematic bias on threshold estimates in the lower part of Figure 3.1. These are estimates based on model 2b-1 with additionally constrained factor means, so that systematic bias is mediated to the thresholds. One can notice an upward shift of all thresholds (cf. formulas (3.9) and (3.11); compare to upper part of Figure 3.1). While there was only non-equivalence on one of the thresholds before, systematic bias now causes systematic non-equivalence on all thresholds (except those constrained as MIC). Note that the strength of the impact of systematic bias somewhat varied across indicators. This was due to the fact that systematic bias was mediated by loadings, reflecting its impact depended on the strength of association between indicator and latent trait.

3.4.5 Duty to Obey the Police scale (DTO)

The DTO scale differed from the prior two in that no explicit DK categories were offered in paper and web. The testing procedure still reproduced nearly all findings from NTP and PV (Table 3.4; note that the configural equivalence model is just-identified and therefore not shown). There was still a difference in systematic bias. Also error variances were consistently smaller for all indicators in web/paper, likewise the NTP and PV questions. However, two differences emerged. First, we found no measurement effects on the scales of any of the indicators (no threshold differences) as suggested by very low RMSEA of model 2, rejecting H2. Second, we found a significant difference

in factor variance of the web condition, while all other variances were equal (model 5-2). This was remarkable, considering differential systematic variance was not found for the NTP and PV scales. In the discussion, we speculate that this is related to the omission of DK in web.

Table 3.4: Equivalence test sequence for the DTO Scale

Model	Equivalence Test	Tested against	RMSEA	RMSEA Diff.	Model Chi²	Adj. Chi² diff. test	Model df (diff.)
Scale Equivalence:							
2	Fixed Scale	-	.028	-	53.6[c]	-	30
Random Error Equivalence:							
3a	Fixed Error Variance	2	.062	+.034[a]	186.3***	157.1***	39 (9)
3a-1	Web=Paper≠F2F=Tel.	2	.032	+.004	71.7***	20.6[c]	36 (6)
Systematic Error Equivalence:							
4a	Fixed Factor Means	2	.039	+.011[a]	82.8***	16.7*[b]	33 (3)
4a-1	Web=Paper≠F2F=Tel.	2	.019	-.009	43.5 (n.s.)	2.4 (n.s.)	32 (2)
5	Fixed Factor Variances	3a-1	.044	+.012 [a]	115.6***	22.8*[b]	36 (3)
5-1	Web=Paper≠Capi=Cati	3a-1	.046	+.014 [a]	117.7***	20.1*[b]	38 (2)
5-2	Paper=Cati=Capi≠Web	3a-1	.030	-.002	72.0***	5.2 (n.s.)	38 (2)

N=3972 (76 cases excluded with Don't know / Refusal on all indicators)
* p<.05, ** p<.01, *** p<.001
'≠' denotes free parameters for two modes, '=' denotes fixed parameters for two modes.
a. Meaningful change of RMSEA criterion (i.e. >.01)
b. Oort adjustment of critical value resulted in a lower significant level. In all other cases adjustment did not change level of significance (Kim & Yoon, 2011; Oort, 1998).
c. Effect / Significance did not hold to cross-validation in both split half samples.

3.4.6 Comparison of factor means and indicator reliability across scales

Two key findings across all scales are different extents of systematic bias and lower random error variances in the self-administered modes. Table 3.5 compares standardized factor mean estimates and reliability estimates across scales. The lower error variance in the self-administered modes manifests in higher indicator reliabilities of most of the questions in paper/web than in F2F/Tel. For the DTO scale it can be seen that higher compound factor variance results in inflated reliabilities (cf. formula (3.7)).

Table 3.5: Mode differences in indicator reliabilities (from models 5, NTP and PV, and 5-2, DTO) and standardized factor means (from models 4a-1 and 4b-1)

	NTP Scale		PV Scale		DTO Scale		
	F2F/Tel	Web/Paper	F2F/Tel	Web/Paper	F2F/Tel	Paper	Web
Indicator Rel. 1	0.590 (.029)	0.678 (.035)	0.545 (.019)	0.645 (.018)	0.317 (.016)	0.396 (.019)	0.490 (.028)
Indicator Rel. 2	0.444 (.029)	0.472 (.028)	0.432 (.019)	0.532 (.022)	0.767 (.023)	0.870 (.021)	0.908 (.017)
Indicator Rel. 3	0.577 (.030)	0.764 (.034)	0.660 (.018)	0.663 (.021)	0.638 (.020)	0.724 (.019)	0.794 (.020)
Indicator Rel. 4	0.101 (.015)	0.118 (.017)	0.771 (.018)	0.835 (.019)	-	-	-
Factor Means[a] (Model 4a-1)	0	.384 (.050)	0	-.323 (.042)	0	-.159 (.041)	-.127 (.033)
Factor Means[b] (Model 4b-1)	0	.192 (.054)	0	-.468 (.059)	-	-	-

a. Model 4a-1: Scale equivalence with free factor means only between F2F/Tel and Mail/Web (cf. **Fehler! Verweisquelle konnte nicht gefunden werden.**)
b. Model 4b-1: Scale non-equivalence between F2F/Tel and Mail/Web (free thresholds) with free factor means only between F2F/Tel and Mail/Web (cf. **Fehler! Verweisquelle konnte nicht gefunden werden.**)
In models 4a-1 and 4b-1, factor variances were additionally equality constrained for Web/Paper to yield equivalent standardized factor means (in the NTP and PV scales only). In the DTO scale, the factor variances were shown to differ strongly between mail and web (model 5-2, Table 3.4). The standardized means differ due to higher factor variances in Web.

Standardized factor means are shown from models 4a-1 and 4b-1 (NTP and PV only). Means of the interviewer modes were constrained to zero for identification. Negative means of the self-administered modes in the NTP scale indicated the systematic bias difference to the interviewer modes. The negative sign suggests that across all categories it was relatively easier in web/paper to answer that a traffic problem persisted more frequently in the neighborhood. For PV and DTO a positive mean was found for web/paper indicating that it was more difficult to agree to questions about proper police visibility in the neighborhood and support of police actions (cf. Figure 3.1). The strength of systematic mean differences varied across scales depending on whether model 4a-1 or 4b-1 is taken as a benchmark. However, the mean difference was smallest for the DTO scale.

3.5 Discussion

Survey researchers designing mixed-mode surveys need to know which modes can safely be combined in later analysis. In the present study, ordinal MCFA models were applied to assess measurement effects of modes on the equivalence of scale, random errors and systematic errors of attitudinal rating scale questions. Consistently with our expectations, we found a divide between interviewer and self-administered modes and nearly complete parity, when comparing F2F with telephone and web with paper. The chief differences between interviewer and self-administered modes were represented by threshold biases, systematic biases, and the extent of random error.

It must again be noted that our study was conducted in the context of a large national survey commissioned by Statistics Netherlands. The consistency of our findings across three scales measuring different traits, using different numbers of answer categories and labels, probability sampling from a general population, as well as the statistical power of our analyses, allow some stronger conclusions about measurement effects of modes. We will discuss implications for survey methodology and statistical modeling of measurement effects separately.

3.5.1 Considerations about survey methodology

The first key finding of the present study is that measurement effects were not indicator-specific phenomena of questions 'sensitive' to mode, but systematically affected all indicators. Foremost, this is caused by unequal extents of systematic bias (H3a confirmed), which by definition affects all thresholds of indicators causing systematic scale non-equivalence. For the NTP and PV scales we could also show that there are individual fluctuations per indicator on thresholds that might be due to question content (H1 and H2 confirmed). On the PV scale only the second threshold of each indicator was affected. This suggests that the indicator-specific threshold bias might rather reflect a systematic threshold-specific bias that cannot be absorbed by the factor means. In this case the indicator-specific effects would not relate to content, but are a second symptom of a systematic category bias.

These results suggest in practice that the same respondent answers the same questions differently, when asked in an interviewer- or a self-administered mode. The

effect was identified across different question topics, formats, and position in the questionnaire. It therefore appears unlikely that these design elements can be altered to mitigate the difference in systematic bias (factor means). Rather the effects are probably caused by mode-specific factors that are impossible to balance by questionnaire design (cf. section 3.1). In survey designs that need to combine more than one survey mode in data collection and analysis, our findings therefore suggest that caution is required when combining data from interviewer and self-administered modes, especially if considerable amounts of attitudinal rating scale questions are to be included. Furthermore, consistent with theoretical expectations, we did not identify any difference in systematic errors nor the item-specific scale parameters between F2F and telephone on the one hand and paper and web on the other (with exception for web on the DTO scale, addressed below). Therefore, the viable mode combinations in mixed-mode surveys appear to be either the interviewer or the self-administered modes. If surveys focus on factual questions, however, implications of the current study generally do not apply, because our study focused on attitudinal questions only. Further research therefore needs to examine measurement effects on factual questions.

The cause of the identified difference in systematic biases between interviewer and self-administered modes might have been stronger social desirable responding in the interviewer modes, if one interprets less frequent reporting of traffic problems, better evaluation of police visibility, and stronger duty to obey police actions all as desirable answers[10]. Reporting neighborhood traffic problems appears to us a topic with least or no social sensitivity, however. Thus it is surprising to identify a difference in systematic bias of equal strength as for evaluations of police visibility (Table 3.5), which arguably has higher sensitivity. We therefore argue that further unknown sources of systematic bias might be present that generalize the problem to non-sensitive rating scale questions. Assessing systematic bias across further scales with a priori low sensitivity is an important aspect for further research. Thereby it is an advantage of the present statistical method that the conclusions drawn above apply regardless of our knowledge of the true cause of difference in systematic biases.

[10] Noteworthy, acquiescence bias is a possible alternative explanation for the systematic bias difference on the PV and DTO scales, but for NTP the bias was in the dis-acquiescent direction (infrequent reporting).

The second key finding of the present study was lower extent of random error and consequently higher reliability of most indicators in the self-administered modes (H1 and H2 confirmed for random error). This finding matches earlier empirical literature on reliability (cf. section 3.1). Noteworthy, the higher reliability is not caused by loading non-equivalence or higher systematic error variance. Only in the web condition of the DTO scale did the increased factor variance further increase reliability estimates. The less pressured situation during self-administration, own pace, time for thought, and the possibility to re-read questions multiple times appear to reduce random impact on measurements. Researchers studying relationships between attitudinal questions can expect less attenuated estimates from web/paper questionnaires. Also, self-administered modes can prove more efficient in the estimation of descriptive statistics.

A further result of this study is that, contrary to systematic bias, no differences in systematic variance were found for NTP and PV scales, and the F2F, telephone and paper modes of the DTO scale (H3b rejected). If prior findings on systematic answer behavior differences, like acquiescence, extremeness, and nondifferentiation, consistently hold across rating scale question, these do not have the expected effect on systematic variance. Possibly, they balance each other in unknown ways or by other factors. Another explanation is the applied propensity score weighting adjustment. Some prior studies did not use this type of adjustment when studying answering behaviors. Also many of the MTMM studies summarized in Saris and Gallhofer (2007) perhaps did not apply nonresponse adjustment. In our unweighted analyses (not shown), there were more pronounced variance differences between interviewer and self-administered modes, which, however, fully disappeared after weighting and after taking the further robustness measures. Systematic variance difference might hence be a selection effect. In general, this is good news, since presence of differential systematic variance would have signified yet another source of systematic scale non-equivalence.

Results on the third scale, DTO, differed from the previous two, but this scale also entailed a major difference in design: web/paper did not present explicit ‘Don’t know’ categories (DK). First, we found increased systematic variance in the web but not in the paper condition. This might be related to the fact that all questions now had to be answered in web due to ‘forced-choice’ administration, while in paper questions could still be skipped. Respondents might have shown straight lining or nondifferentiation

behaviors instead of answering DK in web inflating systematic error variance. An alternative explanation for this finding is, however, the scale's position, which was close to the end of the questionnaire, where web respondents might have shown stronger effects of response burden. Increased systematic variance of web now caused non-equivalence of scale of all indicators between web and paper, which is certainly problematic in face of their strict equivalence in the two other scales.

Second, we did not identify an indicator-specific threshold bias on the DTO scale (scale equivalence). Furthermore, the systematic bias difference on the DTO scale was smaller than on the other scales, though not absent. Even though there are alternative explanations for both findings (e.g. topic, category labeling, and position in questionnaire), omission of DK might thus reduce the systematic bias problem to some extent. Conceptually, this finding can be related to the 'visual scale midpoint', which is biased by DK (Tourangeau et al., 2004). This gives a hint that a part of the systematic bias difference might be caused by the visual presentation of scales and presentation of DK.

In practice, omission of DK might therefore be helpful in reducing the systematic bias problem, while it apparently cannot fully solve it. 'Forced-choice' administration in web probably was counter-productive, but the problem could be solved by allowing respondents to skip questions in web questionnaires without DK option. Assessing the impact of DK on equivalence of systematic bias and variance vis-à-vis alternative explanations that might yield better unified mode designs is an important path for further research.

3.5.2 Considerations about statistical methodology

One decisive advantage of MCFA models over marginal analyses of single questions is the possibility to make inferences about the relative sizes of systematic bias and variance. In the 'compound parameterization' applied in this study, we assumed, consistent with approaches, formulated by Saris and Andrews (1991), Scherpenzeel and Saris (1997), and Alwin (2007, p. 42), that methods, such as modes, invalidate the true score that is estimated, T^*, from the actual true concept of interest, T. An alternative

parameterization for systematic errors would be to specify S as a random effect of unobserved heterogeneity affecting the indicators (cf. formula (3.8)):

$$y^*_{ij|M} = \lambda_{j|M} T_i + S_{i|M} + \varepsilon_{ij|M}. \quad (3.16)$$

This model assumes S_M to be a random effect (or equivalently a factor with unit constrained loading) with a mean and variance. Similar approaches have been suggested in the literature (Billiet & McClendon, 2000; Bollen & Paxton, 1998; Welkenhuysen-Gybels et al., 2003). In Mplus, estimation of this model was possible by specifying a second factor with unit constrained loadings impacting all indicators equivalently. Means and variances of T were zero and unit constrained, respectively, and the model assumed scale equivalence. Subsequently, means and variances of S_M could be tested for equality, representing an alternative equality test of systematic errors, where now S_M is not mediated by factor loadings of T^*. These analyses are illustrated in Appendix B-2. In sum, equality tests about S_M yielded the same conclusions on systematic errors. However, in testing reasons for scale non-equivalence this model caused estimation or identification problems, especially when freeing loadings of T or when leaving variance of S_M free in less constrained models. This turned the approach impractical for the current and other analyses. One possible explanation is that ordinal MCFA models, in which more than one factor loads on the same indicator, have a structure for which exact identification conditions are unknown (Millsap, 2011, p. 130). This parameterization still is an attractive approach to testing systematic error equivalence in the presence of scale equivalence and also to test loading and threshold non-equivalence, provided the estimation problem can be solved. It is both a conceptual and an empirical question, which of the two parameterizations is the 'correct' one. Two of the models relying on the 'compound' parameterization had better fit than the alternative parameterization and one slightly worse. Hence we tend to prefer the compound parameterization on empirical grounds. Conceptually, the question of 'where' modes systematically affect measurement (i.e., on indicator level or on true score level, is a crucial one, but beyond this discussion).

Regardless of particular parameterization, S can crucially affect the equivalence of questions before indicator-specific loading or threshold non-equivalence even need to be considered. Arguably, in face of our findings, it might be less relevant to consider

indicator-specific non-equivalence, if there is a difference in the extent of systematic bias (or variance) to begin with. If a test of systematic bias or variance in the scale equivalence model indicates non-equivalence, the analysis could already stop and conclude about scale non-equivalence of all indicators in a given set of questions. Methods that explicitly model systematic bias or systematic variance, such as MTMM (Saris & Gallhofer, 2007), might thus be viable alternative approaches to studying mode equivalence.

We used ordinal MCFA in our analyses. Alternative categorical measurement models are offered by item response theory. It is noteworthy that IRT models would assume error variance equality (see note 5). As we clearly found error variance differences, we prefer the more flexible MCFA model.

Despite the fact that propensity score weighting on socio-demographic variables was used to adjust for selection bias in all analyses, there still remains a risk that some bias could not be removed successfully, because important confounders might not have been observed. When attempting causal inference using adjustment on background characteristics, experimental mode comparisons always encounter this potential threat to validity. Unfortunately, the assumption that bias is fully removed by the propensity score adjustment typically cannot be tested. An indication in favor of the validity of our results represents, however, the clear consistency of effects conforming to theoretical expectations across scales with three different topics. This reassures that the observed effects are indeed mainly caused by measurement differences between modes.

Whereas adjustment on propensity scores is an established technique to balance selection bias (Morgan & Winship, 2007), it was recently suggested to use instrumental variables in the estimation of measurement effects on the means of survey variables (Vannieuwenhuyze, Loosveldt, & Molenberghs, 2010; Vannieuwenhuyze & Loosveldt, 2013). The instrumental variable method offers great potential in coping with confounding more effectively than statistical approaches that use only background characteristics while possibly omitting important unobserved confounding variables. The technique requires, however, a data collection design that is quite different from the one applied in the present study (in particular, a single mode comparison sample administered in parallel to a mixed-mode sample). Furthermore, the technique, which

originally has been suggested to estimate measurement effects on population means, would first need to be extended for use in more complex variance-covariance models, such as CFA. This is an important path for further research.

The propensity weights were treated as fixed in the analysis, which might underestimate variance, if weights lack precision. Since the maximum normalized weight was small, we believe this threat not to be large. Re-sampling techniques could offer a way to control for the variance of weights. However, to date there is no standard procedure available to combine results of nested hypotheses tests across many sample draws.

Finally, in all analyses the Oort correction and cross-validations were applied. In some cases these robustness measures changed conclusions compared to the unadjusted statistics. These non-robust findings might turn out to have substance, for example, if this study was replicated with larger samples, but the effects would most likely still be small, because the current sample is already quite large. In any case, we advise to apply robustness measures in similar analyses to caution against over-fitting and inflated false positive detection rates

4 A Framework for Estimating Measurement Effects of Survey Modes using Within-Subject Designs [1]

The emergence of new, and rapidly changing, communication channels that use the web and a strong pressure on survey budgets are forcing survey institutes to consider redesigns of their surveys. Along with this trend, research into survey designs that switch traditional survey modes, such as face-to-face, to cost-efficient single-mode designs, such as web, or to multiple modes, so-called mixed-mode designs, is actively pursued (De Leeuw, 2005; Dillman, Phelps, et al., 2009; Vannieuwenhuyze & Loosveldt, 2013).

In taking design decisions of this type, it is a well-known problem that measurement error of questions can differ across modes threatening the accuracy and comparability of data (Krosnick, 1991, 1999; Tourangeau et al., 2000). One option to deal with this problem is to minimize any differences in in measurement error between modes before they occur, for example by means of 'unified mode questionnaires' (Dillman & Christian, 2005; Dillman, Smyth, et al., 2009, p. 326). In designing such questionnaires, estimating the average difference in measurement error between two modes, called the average 'measurement effect' of a question (*ME*, also known as mode effect), is an essential prerequisite.

In the present paper, we discuss how to estimate relevant *MEs* for judging unified mode designs in the two scenarios described above (De Leeuw, 2005; Dillman, Smyth,

[1] This chapter is under review as Klausch, T, Schouten, B., & Hox, J., A Framework for Estimating Measurement Effects of Survey Modes using Within-Subject Designs.

et al., 2009, pp. 306–310; Groves et al., 2010, pp. 175–177): (1) single-to-single mode switches (e.g., during longitudinal or repeated cross-section surveys) and (2) switches to or introduction of sequential mixed-mode surveys.

In scenario (1), cost constraints might necessitate to switch a repeated cross-sectional survey using face-to-face completely to the inexpensive web mode (Biemer, 1988). In this scenario, response distributions of face-to-face often need to be preserved to ensure comparability across time or because face-to-face is known to have small measurement error for a given question. A similar situation is represented by mixed-mode panel surveys that apply different modes in different survey waves (e.g., face-to-face for recruiting respondents and web for re-interviewing; Dillman & Christian, 2005).

In scenario (2), sequential mixed-mode designs, nonrespondents in one mode are followed up by at least one other mode. Often an inexpensive mode, such as web, is used before a more expensive mode, such as face-to-face. Minimizing measurement error is a problem of considerable concern in this design, too. If face-to-face provides accurate answers, for example, the presence of *MEs* implies that web increases measurement error of the mixed-mode survey relative to face-to-face (Dillman, 2009).

Common designs for estimating *MEs* in these scenarios would use a split-ballot (between-subject) mode experiment or a sequential mixed-mode survey. However, since persons self-select for participation in field research even if random sampling is used, estimation of MEs in these designs poses the threat of selection bias. Selection bias occurs if the selection mechanism into modes is not independent of (i.e., it is confounded with) a target variable. A *ME* is then also said to be confounded with a 'selection effect' (Jäckle et al., 2010; Vannieuwenhuyze et al., 2010; Vannieuwenhuyze & Loosveldt, 2013). Unbiased estimation of *MEs* in scenarios (1) and (2) therefore often requires covariate-based adjustment for the confounding problem using techniques such as calibration weighting, regression estimation, or matching on additional auxiliary information (Morgan & Winship, 2007; Schafer & Kang, 2008).

Auxiliary variables need to be exogenous suggesting that they may not underlie *MEs* themselves (Imbens, 2004; Vannieuwenhuyze & Loosveldt, 2013). However, many variables observed in split-ballot or mixed-mode designs may suffer from *MEs*.

For this reason, survey practice has often used socio-demographics as adjustment covariates assuming that socio-demographics do not encounter *MEs* or are available from external sources (Jäckle et al., 2010; Lugtig, Lensvelt-Mulderts, Frefrichs, & Greven, 2011; Schonlau et al., 2004; Tourangeau & Smith, 1996; Vannieuwenhuyze & Loosveldt, 2013). However, socio-demographics are often only weakly related to response mechanisms or target variables. Therefore, it is possible that many *ME* estimates are still biased after adjustment.

Novel to the present paper is that it is suggested to estimate relevant *MEs* in both scenarios by means of auxiliary data collected in so-called 'within-subject' (cross-over) designs instead of using between-subject or sequential mixed-mode designs (Winship & Morgan, 1999). The central advantage of within-subject designs is that they offer stronger adjustment covariates than the alternative options. In a within-subject design, the same sample is approached at two subsequent points in time by two different modes while posing target variables repeatedly. It is then more plausible to adjust for selection bias, because the repeated target variables are strongly related and thus can act as strong auxiliary information.

In within-subject experiments a selection of candidate modes (i.e., modes considered in scenarios 1 or 2) is administered to all relevant population domains. This procedure enables researchers to evaluate potential *MEs* before design decisions are taken and gives questionnaire designers an opportunity to adapt questionnaires (e.g., using unimode principles). Schouten et al. (2013) were one of the first to conduct a type of within-subject experiment (cf. also Biemer, 2001). The authors report on a between-subject mode experiment (face-to-face, telephone, mail, and web) in the context of the Crime Victimization Survey (CVS), where in a mixed telephone - face-to-face follow-up of all subjects, sub-sets of CVS questions were repeated. It was then proposed to decompose overall mode effects into measurement and selection effects by calibrating the mode-specific response samples to the second wave using the re-interview variables for weighting.

The present paper develops a formal framework for estimating *MEs* by within-subject designs of this type. In doing so, we extend the work by Schouten et al. (2013) in some important respects. Our framework distinguishes multiple types of *MEs* that are

applicable in the scenarios (1) and (2) for unified questionnaire design. We describe how to estimate these *MEs* from within-subject data and the required assumptions. In doing so, we differentiate two different types of within-designs: the simple within-subject design (one mode with follow-up) and the within-subject design with control group. In the control group, the same mode is used at both occasions. Since the follow-up is conducted at a later point in time, time-related change of response variables and response mechanisms may impact unbiased inference. The control group may be used to estimate and adjust for such change. Further important assumptions of the within-designs are so-called measurement and response equivalence across occasions, which can be tested in control group designs.

This paper is structured as follows. We first introduce different types of *MEs* and explain how they can be applied for unified mode design in scenarios (1) and (2) (section 4.1). Next, we review estimation of *MEs* based on between-subject designs or sequential mixed-mode data and exemplify the practical limitations when available adjustment variables, e.g. socio-demographics, are weak (section 4.2). Subsequently, we explain how to estimate *MEs* in within-subject designs and extend the design for a control group (section 4.3). Finally, we illustrate the method using a series of practical examples from a large-scale within-subject experiment within the Crime Victimization Survey (section 4.4) re-analysing the data collected by Schouten et al. (2013) in the CVS.

4.1 Types of Measurement Effects and their Utility in Mixed-Mode Design

In this section, we define different types of measurement effects and explain how these can be employed in scenarios (1) and (2) to improve mixed-mode questionnaire design. For simplicity we distinguish modes A and B in the following. In scenario (1), the single-to-single mode switch, we consider the situation when mode A (e.g., face-to-face) is replaced by mode B (e.g., web). In scenario (2), sequential mixed-mode, we distinguish two sub-scenarios. In scenario (2a), we consider the situations when, nonrespondents in mode B (e.g., web) are re-approached by mode A (e.g., face-to-face). In addition we consider the reverse design option, where mode A is followed by B (scenario 2b). It is possible to regard (2a) and (2b) as a situation when mode A is

replaced by sequential mixed-mode or as introduction of a new mixed-mode survey without any particular switch in mind.

We introduce the following notation (cf. Vannieuwenhuyze et al. 2010). Let Y^m denote a continuous or discrete random variable for the outcome on a given question posed under mode M. The average 'marginal ME' is then defined as difference in expectations[2]

$$ME \coloneqq E(Y^b) - E(Y^a). \tag{4.1}$$

The deviations of the mean of outcomes $E(Y^m)$ from their 'true mean' can be defined as 'measurement error'. Therefore, an *ME* of size zero indicates that a question evokes the same extent of measurement error under different modes. This idea becomes central when applying *ME* estimates for unified mode design. Now consider the situation when a survey is administered using a particular mode. In this situation, the fieldwork differences between modes evoke specific 'response mechanisms', denoted by binary response variable S^m, where '$S^m = 1$' indicates response and '0' nonresponse[3].

In principle, measurement error might differ between respondents with higher or lower 'response probability' $P(S^m = 1)$ (Fricker & Tourangeau, 2010; Kaminska, McCutcheon, & Billiet, 2010). Therefore, it makes sense to consider *MEs* also conditional on response in modes A or B. In this case, we call a *ME* 'conditional' (Vannieuwenhuyze & Loosveldt, 2013). There are two conditional *MEs* for respondents:

$$ME^a \coloneqq E(Y^b|S^a = 1) - E(Y^a|S^a = 1) \tag{4.2}$$

$$ME^b \coloneqq E(Y^b|S^b = 1) - E(Y^a|S^b = 1). \tag{4.3}$$

For example, ME^a indicates the difference in measurement error (answers) between modes A and B that can be expected from respondents in mode A. Questionnaire designers often try to minimize absolute measurement error of a survey design against

[2] If Y is discrete, let ME denote the deviations from marginal homogeneity in a category y of the contingency table of Y^b and Y^a, i.e. $ME = E(Y^b = y) - E(Y^a = y)$.

[3] S^m can subsume the outcome of all possible reasons for mode-specific non-observation against the full population, such as refusal, non-contact or non-coverage (Groves, 1989; Groves et al., 2010).

the population value. Since any mode may measure *Y* with error, the presence of a *ME* indicates in most situations that one of the modes measures *Y* with more error than the other for a given group of respondents (2) and/or (3), or marginally (1) in the population. Furthermore, if it can be plausibly assumed, which mode evokes less measurement error on *Y*, questionnaires can be optimized towards this 'benchmark mode'. One way to do so is by unified mode questionnaires, which are designed to evoke equal measurement error from two given modes (De Leeuw, 2005; Dillman & Christian, 2005; Dillman, Smyth, et al., 2009, p. 326) .

An optimal unified mode questionnaire minimizes *MEs* and also maintains measurement error of the mode that is set as benchmark. Now, three objectives may apply: mode A is taken as the benchmark mode, mode B is taken as the benchmark mode or it is unknown what mode produces least error and should act as benchmark. The two scenarios and the three objectives lead to nine situations (Table 4.1).

Table 4.1: Overview on relevant conditional and double-conditional *MEs* for testing zero-constraint hypotheses to evaluate unified mode designs

	Measurement Benchmark Mode		
Design Scenario	Mode A	Mode B	Unified design only (i.e., A or B)
1. Single: A switched to B	ME^b	(always provided)	ME^b
2a. Sequential: B followed by A	ME^b	$ME^{b,a}$	$ME^b \wedge ME^{b,a}$
2b. Sequential: A followed by B	$ME^{a,b}$	ME^a	$ME^a \wedge ME^{a,b}$

When it is the objective to nullify *MEs*, zero-constraint hypotheses on different *MEs* are assessed. First, consider the single-to-single mode switch (scenario 1). Suppose it is the goal to optimize measurement error at the level of mode A. Then the relevant test is represented by

$$H_0\colon\ ME^b = 0.$$

This test is equivalent to evaluating whether $E(Y^b|S^b = 1) = E(Y^a|S^b = 1)$, so that respondents in the 'new' mode B in expectation provide answers that are equal to mode

A (first row, second column). Thus, a unified design would be obtained for this question.

If $ME^b \neq 0$, the question's wording and format need to be reconsidered in order to find a better unified mode design, where, obviously, this might not always be possible in practice. Column three lists the second objective, when mode B is the benchmark. In this case, the 'new' survey using mode B obviously evokes benchmark answers, so that separate testing is not needed. The test in column four ('unified design only') considers the third objective. In some practical cases, the researcher cannot be certain about the size of measurement error in different modes. In this situation one has to act as if either mode A or mode B is the benchmark. This situation thus requires testing $ME^b = 0$ as well.

Second, consider the two sequential mixed-mode designs starting with 2a. The mean of the outcomes provided by respondents at the first step of the sequential design, $E(Y^b|S^b = 1)$, exhibits measurement error of mode B. If mode A is the desired benchmark for the mixed-mode sample (second column, Table 4.1), it needs to be evaluated whether $ME^b = 0$, likewise argued for design 1. This is sufficient, because respondents at the second step of the sequential design already provide benchmark answers. If mode B is the benchmark (third column), the respondents at the second occasion provide answers with measurement error of mode A which might differ from mode B. The mean outcome of this group is $E(Y^a|S^b = 0, S^a = 1)$. Since answers in this follow-up group should not differ from mode B in expectation, we require

$$ME^{b,a} \coloneqq E(Y^b|S^b = 0, S^a = 1) - E(Y^a|S^b = 0, S^a = 1) = 0. \quad (4.4)$$

$ME^{b,a}$ is called a 'double-conditional *ME*', because it conditions on two response mechanisms. It applies to a sequential mixed-mode design, which follows up nonrespondents in mode B by mode A. The hypotheses for design 2b follow the same logic as for 2a and can be taken from Table 4.1 (last row). However, now the double conditional *ME* is defined as:

$$ME^{a,b} \coloneqq E(Y^b|S^b = 1, S^a = 0) - E(Y^a|S^b = 1, S^a = 0) \quad (4.5)$$

for a sequential design following up mode A by mode B at occasion two. Under the third objective, both, the conditional and double-conditional effects need to be assessed simultaneously to guarantee a unified design.

In conclusion, it is important to realize that marginal *MEs* defined in (1) are irrelevant in all scenarios. Estimation can, therefore, focus only on (double-) conditional instead of marginal *MEs*, formulas (4.2) to (4.5). The next section discusses the available approaches and assumptions needed for estimating the defined quantities before section 4.3 presents the within-subject methods.

4.2 Estimation in Between-Subject and Sequential Mixed-Mode Designs

The most common approach to estimating conditional (and marginal) *MEs* are experiments using 'between-subject' (split-ballot) designs (Figure 4.1), for which separate independent samples are drawn and assigned to different modes (Aquilino, 1994; Chang & Krosnick, 2009; Dillman, Phelps, et al., 2009; Fricker et al., 2005; Heerwegh & Loosveldt, 2008; Heerwegh, 2009; Jäckle et al., 2010; Kreuter et al., 2008; Schonlau et al., 2004; Tourangeau & Smith, 1996). An alternative to these designs is using a mixed-mode survey directly (Vannieuwenhuyze et al., 2010; Vannieuwenhuyze & Loosveldt, 2013). This approach is useful when scenario 2 is of interest or data from an ongoing mixed-mode survey are readily available.

The key problem in both designs is the presence of missing data, of which two different types are distinguished in Figure 4.1. The left side shows the missing data pattern of a between-subject design. Depicted are the variables Y^a and Y^b introduced in section 4.1 and auxiliary information *X*. The white areas represent the observed part of the data (i.e., response, $S^a = 1$ and $S^b = 1$) and the black areas missing data due to unit nonresponse ($S^a = 0$ and $S^b = 0$), the first type of missing data.

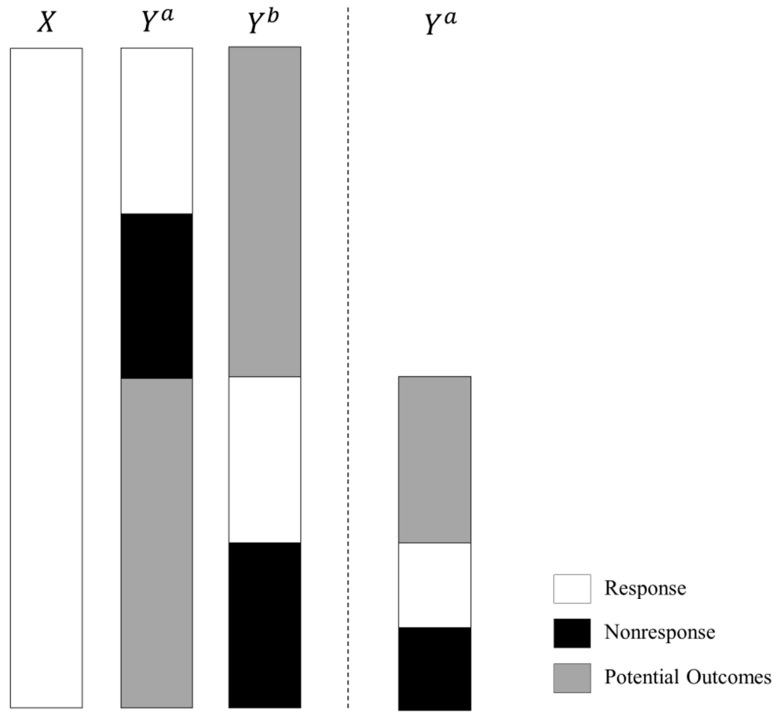


Figure 4.1: Missing data pattern of a Between-Subject Design (left) and a sequential mixed-mode follow-up to mode B in mode A (right)

Furthermore, the moment in time Y^a and Y^b have been observed under a given mode all outcomes in other modes are considered 'potential' (Holland, 1986; Rubin, 1974, 1976, 1977, 2005). Potential outcomes can never be observed in reality, because it is not possible to observe outcomes under two modes at the same point in time. Therefore, the potential outcomes represent the second type of missing data (grey colour). *X* is exogenous auxiliary information (i.e., unaffected by any *MEs*) and either available for all respondents or for all units (Imbens, 2004). In the latter situation *X* is usually accessible from external sources, such as registers. In the former situation *X* is observed as a survey variable and assumed to be exogenous.

The missing data pattern of the follow-up step of a sequential mixed-mode design is depicted on the right (Figure 4.1), where mode A follows B (scenario 2a). It can be seen that potential outcomes are created by not following up respondents in mode B, whereas some response under A is obtained from nonrespondents in B.

A first approach to estimate *MEs* is represented by taking the simple difference in means between response samples. We call this estimator 'naïve', because it is only

equal to marginal or conditional *MEs* under special circumstances. Otherwise, it is said to suffer from selection bias. The expected value of the naïve estimator in the between-subject design is

$$E\left(\widehat{ME}_{BS}^{naive}\right) = E(Y^b|S^b = 1) - E(Y^a|S^a = 1). \tag{4.6}$$

The selection bias of the marginal *ME* is given by

$$SE = \Big(E(Y^b|S^b = 1) - E(Y^b)\Big) - \big(E(Y^a|S^a = 1) - E(Y^a)\big). \tag{4.7}$$

Furthermore, when using $\widehat{ME}_{BS}^{naive}$ as an estimator for ME^b, the selection bias is

$$SE(Y^a) = E(Y^a|S^b = 1) - E(Y^a|S^a = 1) \tag{4.8}$$

and for ME^a it is

$$SE(Y^b) = E(Y^b|S^b = 1) - E(Y^b|S^a = 1). \tag{4.9}$$

It can be shown that all selection biases denote a difference in 'selection error', if these errors are defined as the difference in response sample mean of Y^a and Y^b and their population means. The naïve estimator is only unbiased, if selection bias is negligibly small, suggesting that selection errors are equal across modes (or zero). In this context, selection biases are also referred to as 'selection effects' (Vannieuwenhuyze & Loosveldt, 2013). Thus, $\widehat{ME}_{BS}^{naive}$ is an estimator for an 'overall mode effect' in the between-subject design, confounding selection bias (selection effects) with *MEs*, e.g. for conditional *MEs*

$$E\left(\widehat{ME}_{BS}^{naive}\right) = SE(Y^a) + ME^b = SE(Y^b) + ME^a. \tag{4.10}$$

If only the sequential mixed-mode data is available, the naïve *ME* estimator may be defined differently, as the difference in response sample means of mode B and follow-

up A. Frankly, this estimator may suffer from selection bias as well, but its biases differ[4] from $\widehat{ME}_{BS}^{naive}$.

If selection biases are not negligible, any technique leading to unbiased estimates of conditional or marginal *MEs* 'disentangles' *MEs* from their selection biases. Given the missing data problem illustrated in Figure 4.1, missing data adjustment is the primary approach to unbiased *ME* estimation. Such techniques exploit the auxiliary exogenous data in *X*. Various techniques have been discussed in the context of the causal inference literature, focusing on the estimation of potential outcomes (Imbens, 2004; Kang & Schafer, 2007; Rubin, 2005; Schafer & Kang, 2008) and nonresponse in sample surveys focusing on the adjustment of selection (or nonresponse) error (Särndal & Lundström, 2005). Techniques include calibration weighting, (robust) regression estimation (Bethlehem, 1988, 2002; Cochran, 1977; Kang & Schafer, 2007; Särndal & Lundström, 2005), matching (Rosenbaum, 2002), and multiple imputation (Rubin, 1987; Schafer, 1997).

When estimating marginal *MEs* using the between-subject design, all of these techniques are based on the assumption that nonresponse in both samples is missing at random (MAR) given auxiliary information *X* implying (Little & Rubin, 2002; Rubin, 1976)

$$Y^m \perp S^m \mid X \tag{A1}$$

where $\perp$ denotes independence of Y^m and S^m (here conditional on *X*). (A1) suggests that selection error on Y^m is fully explained by *X*. *X* needs to be available for all units for this purpose.

To the contrary, for estimating the (double-) conditional *MEs*, the conditional means of potential outcomes, i.e. $E(Y^a|S^b = 1)$ and $E(Y^b|S^a = 1)$, are needed. It is never possible to observe these quantities in reality, but they can be estimated assuming MAR as

[4] The expectation of the naïve estimator in the mixed-mode case is $E\left(\widehat{M}\mathrm{E}_{MM}^{naive}\right) = E(Y^b|S^b = 1) - E(Y^a|S^b = 0, S^a = 1)$. Its selection bias against ME^b is, for example, $E(Y^a|S^b = 0, S^a = 1) - E(Y^a|S^b = 1)$, denoting the difference in selection error on Y^a between initial response sample in mode B and follow-up sample (a selection effect).

$$Y^a, Y^b \perp S^a \mid X, S^b = 1 \quad \text{(A2a)}$$

$$Y^a, Y^b \perp S^b \mid X, S^a = 1. \quad \text{(A2b)}$$

These assumptions are also known as 'unconfoundedness' or 'ignorable treatment assignment' in the causal inference literature[5] (Imbens, 2004; Kang & Schafer, 2007; Schafer & Kang, 2008). A2a/b imply that *X* explains the distributional differences between response samples on Y^a and Y^b in the between-subject or mixed-mode design.

For example, in the between-subject design a regression estimator of $E(Y^a|S^b = 1)$ can be constructed as follows. First, a regression model of Y^a on *X* is fitted using the observed data in mode A, $E(Y^a|X, S^a = 1) = X\beta$. Second, potential outcomes $E(Y^a|X, S^b = 1)$ are predicted from estimated parameters $\hat{\beta}$. Assuming A2a/b, we have $E(Y^a|X, S^a = 1) = E(Y^a|X, S^b = 1)$. The mean of predicted potential outcomes over *X* therefore serves as an unbiased estimator of $E(Y^a|S^b = 1)$. A similar regression estimator can be used when mixed-mode data are available[6].

Assumptions (A1) and (A2a/b) are more likely to hold in practice, when *X* strongly relates to the response mechanism and response variables. Usually, socio-demographic variables are available as auxiliary data (*X*) in survey research. Socio-demographics have been applied as 'control covariates' in regression models (Heerwegh & Loosveldt, 2008; Tourangeau & Smith, 1996), as weighting variables (Chang & Krosnick, 2009; Fricker et al., 2005; Holbrook et al., 2003; Jäckle et al., 2010; Klausch, Hox, & Schouten, 2013b; Schonlau et al., 2004), or for matching (Lugtig et al., 2011). Survey practice has shown, however, that socio-demographics are seldom strongly related to

[5] It can be shown that MAR assumptions, when conditioned on a second response mechanism as in A2a/b, are equivalent to the unconfoundedness (or ignorable treatment assignment) assumption typically made in causal inference theory (Rubin, 1974; Imbens, 2004). In this literature, 'treatment assignment' is normally indicated by one selection mechanism only, say $M = \{a, b\}$. In the context of a between-subject design, this notation suggests that 'response is through mode A or mode B', limiting the parameter space to the group of respondents in either mode (excluding nonrespondents). Our notation resorts to more than one selection mechanism in order to be able to describe mode-specific unit nonresponse. Unconfoundedness requires that $Y^a, Y^b \perp M \mid X$ which is equivalent to A2a/b in the between-subject design.

[6] It should be noted however, that in a sequential mixed-mode design where mode B is followed up by mode A (scenario 2a), only ME^b and $ME^{b,a}$ but not ME^a and $ME^{a,b}$ can be estimated. Considering Table 1, these MEs are sufficient for judging unified mode designs in scenario 2a.

response mechanisms or many survey target variables (Couper et al., 2007; Nicoletti & Peracchi, 2005) and differences in socio-demographic distributions between modes are often small (Klausch et al., 2013a). Adjusted estimates also often do not differ greatly from unadjusted estimates (Schonlau, van Soest, Kapteyn, & Couper, 2009). It is, therefore, possible that assumptions (A1) or (A2a/b) do not hold and *ME* estimates may still be biased after adjusting for socio-demographic differences between response samples. Alternative approaches to missing data adjustment can apply so-called 'front-door'[7] or 'instrumental'[8] variables (Vannieuwenhuyze et al., 2010, 2014). As discussed in the next section, the within-subject method relies on missing data adjustment, but use stronger auxiliary information.

4.3 Estimation in Within-Subject Designs

In essence, the within-subject design (WSD) method allows estimating conditional and double-conditional *MEs* needed to take design decisions in scenarios (1) and (2) under weaker MAR assumptions. The method does not allow estimating marginal *MEs*, however. As outlined in section 4.1, this is, fortunately, also not a necessity for effective unimode questionnaire design. We first present the basic outline and necessary assumptions of the simple WSD method and the general principles of estimation. Some assumptions may be tested and relaxed using a control group extension of the WSD. The use of this group is explained subsequently.

---

[7] A ME might be explainable by a mediating factor, the so-called 'frontdoor variable', such as 'survey enjoyment' (Vannieuwenhuyze, Loosveldt, & Molenberghs, 2014). If the frontdoor variable is not affected by selection effects itself and fully mediates the ME between mode and target variable, conditional MEs can be estimated using a method described by Pearl (Morgan & Winship, 2007, pp. 224–230; Pearl, 2009, pp. 81–85). However, currently, there is not any known set of variables that would plausibly fulfil the criteria required from frontdoor variables.

[8] The method supposes scenario 2a where a 'comparison sample' is surveyed additionally in mode A only (Vannieuwenhuyze, Loosveldt, & Molenberghs, 2010; Vannieuwenhuyze & Loosveldt, 2013). To estimate the conditional and the double-conditional ME^b and $ME^{b,a}$, it is then assumed that: $E(Y^a|S^a = 1 \cup S^b = 1) = E(Y^a|S^a = 1)$. This 'representativeness assumption' says that selection error is equivalent for the sample realized by the mixed-mode survey and the comparison survey in mode A. To the contrary, mixed-mode surveys are often regarded as a solution to reduce selection error below that of single-mode designs. However, in this case the assumption would not hold. Despite the theoretical importance of the method, its practical applicability seems somewhat limited for this reason.

4.3.1 Outline and basic assumptions of the simple within-subject design

Contrary to a between-subject design, the within-subject experiment requires surveying the full sample only by a single mode at a first occasion. After some time has elapsed, this sample is approached again at a second occasion, but then in a different mode. For exposition, we assume mode B is presented first and the re-interview is conducted in mode A (Figure 4.2). Presentation of modes in reverse order (A-B) is possible and may be chosen depending on practical considerations. We return to the aspect of order and exchangeability of modes in the discussion.

The goal of the repeated survey is collecting additional information from respondents in mode A. There are two types of variables collected in this re-interview. First, Y_1^b is re-measured as Y_2^a. Indices '1' and '2' denote observations at the respective occasions *t*. As discussed in detail in the following, the repeated measure is then exploited for two purposes: first, for estimating means of *Y* under mode A and, second, as auxiliary data for missing data adjustment. A second option for using the re-interview is collecting any further information that may be related to the response mechanism or target variables referred to as Z_2^a. The primary purpose of *Z* is to additionally improve the missing data adjustment.

A WSD leads to a new missing data pattern (Figure 4.2). Likewise between-subject designs, WSDs also encounter unit nonresponse under both modes. The first occasion provides observations from respondents ($S_1^b = 1$) on Y_1^b and unit nonresponse ($S_1^b = 0$). At the second occasion S_2^a and Y_2^a are observed. This missing data pattern is equivalent for the other variables *Z* (for simplicity, *Z* therefore is not included in Figure 4.2). *X*, likewise in the between-subject design, are variables available for all units[9].

Now, a central step in estimating MEs from WSDs is using Y_2^a and S_2^a as substitutes for the potential outcomes Y_1^a and S_1^a, which are not observed by design at occasion one (grey area, Figure 4.2). This procedure is based on two types of assumptions: time-stability and equivalence of occasions. These assumptions may appear stringent at first.

[9] We make a difference between *X* and *Z* for illustrative purposes. *X* denotes exogenous auxiliary information that is available for all units, for example from an external register. *Z* is surveyed at occasions 1 or 2 as any further information turning the MAR assumption plausible. *Z* therefore shows the same unit nonresponse pattern as Y_1^b and Y_2^a, whereas *X* is available for all units.

We discuss conditions that turn the assumptions more plausible in section 4.3.3 and test and relax them in a control group extension of the design in section 4.3.4.

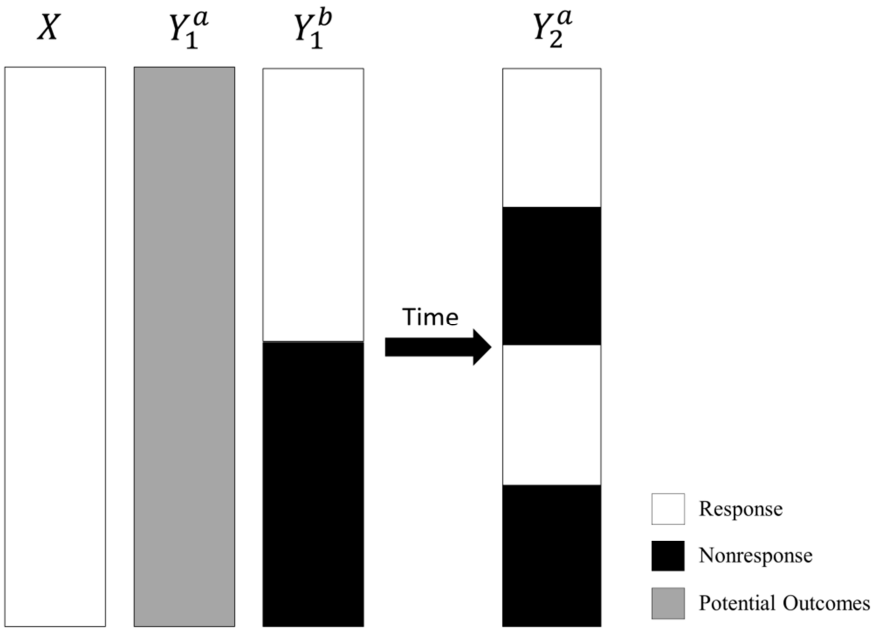


Figure 4.2: Missing data pattern of a Within-Subject Design

4.3.1.1 Time-stability assumptions

Time-stability can relate to, both, the measurements of Y_1^a and the response mechanism S_1^a. First, answer distributions may differ between occasions due to factors related to the progression of time, such as seasonal change of true-scores. In this case Y_2^a is biased against Y_1^a. The means of response distributions are called 'time-stable' if

$$E(Y_2^a|S_1^m = 1) - E(Y_1^a|S_1^m = 1) = 0, \quad \text{(A3)}$$

for any mode *m*. Second, the response mechanism S_1^a may change across time, if respondents have different propensities to participate with respect to Y_t^m. If

$$E(Y_t^m|S_2^a = 1) - E(Y_t^m|S_1^a = 1) = 0, \quad \text{(A4)}$$

for any mode *m* and occasion *t*, a change in response propensities does not affect the response means of Y_t^a or Y_t^b. Selection errors may then be called time-stable.

4.3.1.2 Measurement and response equivalence assumptions

Being a respondent or a nonrespondent at occasion one might also influence the levels of Y_2^a or response probabilities defined by S_2^a. For example, respondents at both occasions might reproduce answers from the first occasion. In this situation, bias of unknown size would be created on Y_2^a. Therefore, we require these quantities to be equivalent in expectation as follows.

Let ${}_1^m Y_2^a$ denote the response distribution given that mode M (mode B in Figure 4.2) was assigned before measuring Y_2^a. The assumption that response distributions in expectation were not impacted by mode assigned at occasion one is called 'measurement equivalence'. Measurement equivalence implies

$$E({}_1^m Y_2^a | S_t^m = 1) - E(Y_2^a | S_t^m = 1) = 0 \quad \forall m, t. \tag{A5}$$

Similarly, the assumption that the mode assigned at occasion one did not impact the expected value of Y_t^m for respondents at the second occasion is called 'response equivalence'. Let ${}_1^m S_2^a$ denote the response mechanism at occasion two given that mode M was assigned at occasion one. Response equivalence implies

$$E(Y_t^m | {}_1^m S_2^a = 1) - E(Y_t^m | S_2^a = 1) = 0 \quad \forall m, t. \tag{A6}$$

It should be noted that both assumptions are weaker than requiring equality in distributions almost surely (i.e., $P({}_1^m Y_2^a = Y_2^a) = 1, P({}_1^m S_2^a = S_2^a) = 1$). This is sufficient, because we focus estimation of means.

4.3.1.3 Missing at random assumptions

Estimation of WSDs is now based on the following two MAR assumptions

$$Y_2^a \perp S_2^a \mid X, Y_1^b, Z_1^b, S_1^b = 1 \textit{ (Forward-directed MAR)} \tag{A7a}$$

$$Y_1^b \perp S_1^b \mid X, Y_2^a, Z_2^a, S_2^a = 1 \textit{ (Backward-directed MAR).} \tag{A7b}$$

(A7a) is called forward-directed MAR, because missing data at a later point in time are ignorable on their earlier measurements (and further information *X* and *Z*; Figure 4.3, left). Similarly, (A7b) is called backward-directed MAR, because earlier missing data

are ignorable on later observations (besides *X* and *Z*, Figure 4.3, right). Both assumptions are extensions of the 'unconfoundedness' assumptions in between-designs, (A2a/b), by information from the within-design. However, since the partial correlation of Y_1^b and Y_2^a is probably strong in most cases, and possibly additional information[10] is available in *Z*, (A7a/b) appear much more plausible than (A2a/b) when using only socio-demographics as auxiliary data. Schouten et al. (2013), for example, measured attitudes about surveys as *Z* besides repeating *Y* in a WSD.

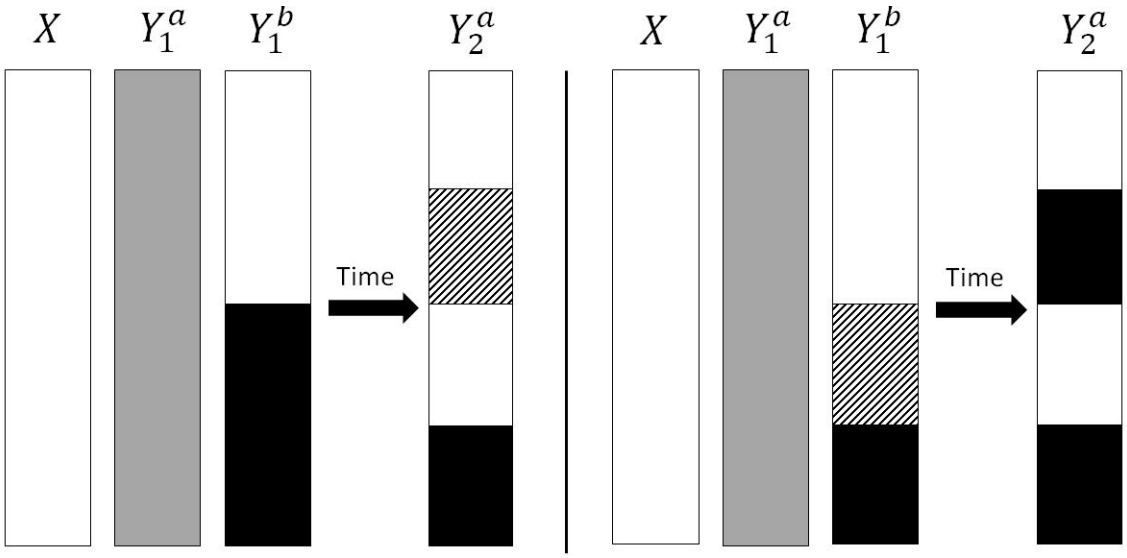


Figure 4.3: Illustration of the ignorable part of missing data (dashed areas) assumed as forward-directed MAR (left) and backward-directed MAR (right)

We do not assume, however, that (A7a/b) would hold marginally, because outcomes are not observed for the group of individuals who are nonrespondents under both modes. Estimating marginal *MEs* would still require ignorable nonresponse on *X* observed for all respondents as stated, for example, in (A1).

[10] As explained above, Z_1^b and Z_2^a denote any other variables surveyed at either occasion one or two that do not represent repeated measures. In designing within-subject experiments, variables that explain differences in response between modes or other survey target variables that are similar to *Y* are useful candidates.

4.3.2 Estimating conditional and double-conditional MEs

In estimating the conditional ME^b and ME^a at occasion one, four means are of interest (formulas (4.2) and (4.3)). For ME^b, we estimate $E\left(Y_1^b|S_1^b=1\right)$ and $E\left(Y_1^a|S_1^b=1\right)$. $E\left(Y_1^b|S_1^b=1\right)$ is estimated directly as response sample mean from the observed data at the first occasion. Since, the mean of potential outcomes $E\left(Y_1^a|S_1^b=1\right)$ is not observable by design, however, we substitute the quantity $E\left(Y_2^a|S_1^b=1\right)$ instead. In doing so, we assume time-stability and measurement equivalence, (A3) and (A5).

Obviously, some information to estimate $E\left(Y_2^a|S_1^b=1\right)$ is missing due to unit nonresponse at occasion two (Figure 4.3, left). Assuming forward-directed MAR (A7a), the quantity can be estimated, however, using techniques such as regression estimation, weighting, matching, or imputation (for an example of regression estimation, see section 4.2). We call this approach the forward method.

For ME^a, we estimate $E(Y_1^a|S_1^a=1)$ and $E\left(Y_1^b|S_1^a=1\right)$. Contrary to ME^b, this step requires substituting both quantities by elements from the re-interview. First, we estimate $E(Y_2^a|S_2^a=1)$ as substitute of $E(Y_1^a|S_1^a=1)$. Under time-stability, (A3) and (A4), and measurement and response equivalence, (A5) and (A6), these means are equal. Second, estimating the mean $E\left(Y_1^b|S_1^a=1\right)$ exploits answers Y_1^b from respondents at the first occasion under mode B, but the answers of respondents to mode A are actually needed. The within-design facilitates observing the response mechanism A at the second occasion, so that we estimate the conditional mean $E\left(Y_1^b|S_2^a=1\right)$ instead. Both means are equal assuming (A4) and (A6).

Again a missing data problem is created by nonrespondents under mode B at occasion one, who are respondents at occasion two (Figure 4.3, right). Assuming data are backward-directed MAR (A7b), the conditional mean can be estimated. We call this approach the backward method.

Double-conditional *MEs* are defined for sequential mixed-mode designs (cf. section 4.1), and condition on two response mechanisms. In particular, $ME^{b,a}$ (formula (4.4)) is defined for respondents to mode A who are nonrespondents under mode B. The mean $E\left(Y_2^a|S_1^b=0,S_2^a=1\right)$ is observed in the WSD. In fact, sequential mixed-mode designs

can be regarded as a variant of WSDs, in which respondents at the first occasion are not followed up in an alternative mode.

Assuming time-stability in the sub-set defined by $\{S_1^b = 0, S_2^a = 1\}$, we now only need to estimate $E(Y_1^b | S_1^b = 0, S_2^a = 1)$, for which observations are missing due to nonresponse under mode B at occasion one[11]. Assuming backward-directed MAR, (A7b), this mean can be estimated.

The second double-conditional $ME^{a,b}$ exchanges the order of modes. Now nonrespondents under mode A, who are respondents under mode B are of interest. Since equivalence across occasions is assumed, the order in which modes are presented is irrelevant for $ME^{a,b}$. Estimation then assumes forward-directed MAR (A7a).

4.3.3 Design considerations about time-stability and equivalence

Clearly, estimating MEs from between- and within-subject designs is based on differing assumptions (Table 4.2). Whereas the between-design (BSD) assumes unconfoundedness (A2a/b) given exogenous *X*, estimating ME^b is possible in within-designs (WSD) under a weaker MAR assumption at the expense of assuming time-stable measurements and measurement equivalence, (A3) and (A5). Estimating ME^a additionally assumes response stability and response equivalence, (A4) and (A6), mainly due to the fact that $E(Y_2^a | S_2^a = 1)$ needs to substitute $E(Y_1^a | S_1^a = 1)$. Based on this conjecture, it is crucial to evaluate under which circumstances the assumptions are more likely to hold in practice, before we relax some of them using a control group extension.

First, the final design scenario for which *MEs* are estimated should be considered (Table 4.1). If scenario 2 is of interest (sequential mixed-mode) the within-design fieldwork period may be designed similarly to the final survey. The time effects and non-equivalence may then be regarded as part of the sequential design, which are generally considered ignorable in practical research. If scenario 1 is of interest or both

[11] It should be noted that (A3) and (A4) only define time-stability in the sub-set of $\{S_t^m = 1\}$. Estimation of $E(Y_1^a | S_1^b = 0, S_1^a = 1)$ formally requires an extension of these time-stability assumptions to the set $\{S_1^b = 0, S_2^a = 1\}$.

scenarios are considered as possible alternatives, however, further considerations about time-stability and equivalence are necessary.

Time-stability depends on exogenous factors related to the progression of time, such as seasonal change of true scores (A3) or survey climates in societies (A4). The extent of change thus also depends on the type of target variables and populations at hand. In general, however, shorter time lags between occasions let time-stability appear more plausible regardless of particular variables and populations.

Table 4.2: Overview on assumptions for estimating (double-) conditional MEs in between-subject, within-subject, and within-subject control-group designs

			Forward method $(ME^b, ME^{b,a})$		Backward method $(ME^a, ME^{a,b})$	
	Assumption	BSD	WSD	WSCGD	WSD	WSCGD
A1	Marginal MAR on *X*	-	-	-	-	-
A2a/b	Unconfoundedness on *X*	X	-	-	-	-
A3	Time-stable *Y*	-	X	-	X	-
A4	Time-stable selection error	-	-	-	X	-
A5	Measurement equivalence	-	X	X	X	-
A6	Response equivalence	-	-	-	X	X
A7a/b	Forward/Backward MAR	-	X	X	X	X
A8a/b	Forward/Backward MAR in control	-	-	X	-	X
A9	Time-stable selection effects	-	-	X	-	-
A10	Time-stable measurement effects	-	-	-	-	X

BSD: Between-Subject Design; WSD: Within-Subject Design; WSCGD: Within-Subject Control-Group Design

Measurement and response equivalence depend on design inherent factors and thus might be controllable to certain extents by design. Measurement equivalence (A5) may depend on two factors. On the one hand, respondents' ability and motivation to reproduce answers from the first occasion affects equivalence. Cognitively salient questions take longer time to be forgotten and thus require longer time lags between occasions. On the other hand, the structure of the follow-up questionnaire may change

the extent of measurement error on Y_2^a. If the follow-up questionnaire differs to large extents from occasion one, a substitution of Y_1^a by Y_2^a may become invalid.

Response equivalence (A6) may be violated due to heavy response burden during the first occasion. Several fieldwork precautions can reduce this risk, however. Interview length of the first survey should be short. During the recruitment of the second survey it should be made clear to all individuals why (repeated) participation is necessary (e.g., to ask additional questions on same topic). If interviewers are employed at the second occasion, they require special training to deal with respondents' questions about the first occasion. Furthermore, in order not to influence response probabilities at the first occasion, individuals should be kept unaware at occasion one about the follow-up survey[12]. Finally, longer time lags between occasions also turn response equivalence more likely, as response probabilities can normalize across time.

Curiously, time-stability is more likely to hold for shorter, whereas both equivalence assumptions are more likely for longer time lags. Hence, there is a trade-off in terms of the timing of occasions. We suggest that prior knowledge about target variables and response probabilities (e.g., time-series data) can give an additional indication about the time frame in which stability can be assumed. If time-stability can be assured, choosing longer time lags is always advisable. Another option is pre-study that evaluates across which time periods important target variables and response probabilities can be considered time stable and equivalent.

Not always can a thoroughly designed WSD reduce the risk that time stability or equivalence do not hold so that estimates of the conditional *MEs* get biased. Therefore, separate testing of the assumptions using a control group offers additional reassurance, as discussed in the next section.

4.3.4 Testing and adjusting for time-instability and non-equivalence

Control groups represent a common approach to exclude time-related change as an alternative explanation to a treatment in interrupted time-series designs (Winship &

[12] Note that earlier we implicitly assumed that the response and measurement at occasion one is not affected by the introduction of occasion two later on.

Morgan, 1999). A control group is created by sampling an independent, parallel sample, in which the mode is kept constant across occasions to avoid confounding of *MEs* with time-related change. The resulting design is called a 'within-subject control-group design' (WSCGD).

If mode A is used for the control group, WSCGDs are equivalent to between-subject designs extended for a second measurement occasion (Figure 4.4). Based on this design it is possible to estimate and adjust for time-related bias due to violation of (A3) or (A4), assuming that change observed in the control group is equivalent to the sample initially interviewed in mode B ('treatment'). In addition, the equivalence assumptions, (A5) and (A6), can be tested to some degree.

For both purposes, various mean components that are subject to nonresponse are estimated from the control group. Therefore, we extend MAR assumptions (A7a/b) in the WSD to the control group

$$Y_2^a \perp S_2^a \mid X, Y_1^a, Z_1^a, S_1^a = 1 \textit{ (Forward-directed MAR in control group)} \quad \text{(A8a)}$$

$$Y_1^a \perp S_1^a \mid X, Y_2^a, Z_2^a, S_2^a = 1 \textit{ (Backward-directed MAR in control group).} \quad \text{(A8b)}$$

The following steps lead to adjusted estimates of ME^b and ME^a. It should be noted that adjustment of the double-conditional *MEs* is not straight forward in WSCGDS. As noted above, if double-conditional *MEs* are estimated for sequential mixed-mode designs, the fieldwork period of the WSD should reflect the final sequential design, so that time-stability is plausible and non-equivalence of response may be considered part of the design.

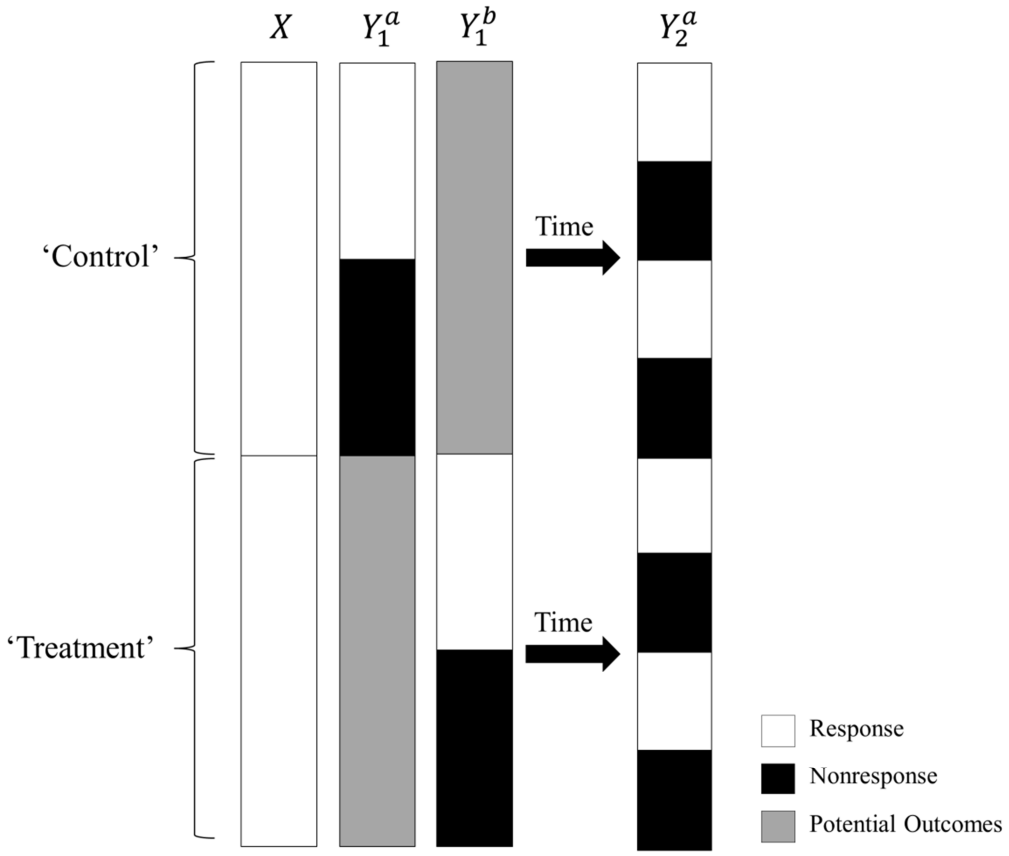


Figure 4.4: Missing Data Pattern of a Within-Subject Control-Group Design

4.3.4.1 Testing and adjusting time-related bias of ME^b

The forward method might introduce bias in the estimator of ME^b, if Y_1^a changes across time in the group of respondents at occasion 1, violating (A3), i.e.

$$\Delta_1 \coloneqq E(Y_1^a|S_1^b = 1) - E(Y_2^a|S_1^b = 1), \tag{4.11}$$

because observations from the second occasion are used instead of the potential outcomes at the first (cf. Table 4.2). Data from the control group design is used to estimate this bias, assess its size, and adjust the estimator appropriately.

However, whereas the control group design can be applied to estimate change in Y_1^a for respondents in mode A, Δ_1 is defined for respondents in mode B. It, therefore, needs to be assumed that change on Y_1^a is independent of (i.e., does not interact with) responding in modes A or B

$$\Delta_1 = E(Y_1^a|S_1^a = 1) - E(Y_2^a|S_1^a = 1). \quad \text{(A9)}$$

The assumption implies that selection effects on Y_t^a do not depend on the time-point that Y_t^a is observed ('time-stable selection effects'). Whereas this is still an assumption, it is only violated, if there are 'second order effects' on the change. We allow for a change in Y_t^a for the group of respondents, but it does not matter which response group (A or B) is used to estimate this change. This can be considered a weaker assumption than assuming time-stability for Y_t^a overall for respondents in mode B. Therefore, (A3) is relaxed. Assuming (A8a), the mean components of Δ_1 can be estimated.

The implication of this adjustment step may be taken from Table 4.2 (column 'WSCGD'). Whereas time-stable measurement (A3) is assumed in the WSD, the WSCGD relaxes this assumption by making weaker assumption (A9).

4.3.4.2 Testing and adjusting time-related bias of ME^a

Estimating ME^a can lead to two types of time-related biases in the WSD. First, we have bias

$$\Delta_2 \coloneqq E(Y_1^a|S_1^a = 1) - E(Y_2^a|S_2^a = 1), \quad \text{(4.12)}$$

caused by substituting the response mean of Y_1^a at occasion one by observations from occasion two. This step requires, both, time-stability of measurement and response[13], (A3) and (A4), besides the equivalence assumptions, under which $\Delta_2 = 0$. In the control-group, $E(Y_1^a|S_1^a = 1)$ is observed, however. Therefore, this bias is avoided immediately and the assumptions are relaxed. This conjecture may be considered a strong feature of the control group design in estimating ME^a.

Second, we might introduce bias

$$\Delta_3 \coloneqq E\left(Y_1^b\middle|S_2^a = 1\right) - E\left(Y_1^b\middle|S_1^a = 1\right), \quad \text{(4.13)}$$

[13] Note that $\Delta_2 = \Delta_1 + E(Y_2^a|S_1^a = 1) - E(Y_2^a|S_2^a = 1)$. Under (A3) $\Delta_1 = 0$ and under (A4) $E(Y_2^a|S_1^a = 1) - E(Y_2^a|S_2^a = 1) = 0$.

which is caused by exploiting the response mechanism at the second occasion instead of the first (A4). This bias is equivalent to the change in selection error on Y_1^b from occasion one to occasion two.

Using the control group it is again possible to estimate this bias, when we assume that the change in selection error on Y_1^b is equivalent to the change on Y_1^a, which is observable in the control group

$$\Delta_3 = E(Y_1^a|S_2^a = 1) - E(Y_1^a|S_1^a = 1). \quad \text{(A10)}$$

Thus, we allow for a change in $E\left(Y_1^b\middle|S_t^a = 1\right)$ over *t*, but we assume that it is irrelevant, whether measurements of mode A or B are used to estimate this change. The assumption implies that ME^a does not depend on the time point on which the selection mechanism S_t^a of mode A is observed ('time-stable conditioning of measurement effects'). Therefore, (A4) is relaxed. Assuming (A8b), the mean components of Δ_3 can be estimated.

In contrast to the WSD, the WSCGD relaxes measurement and response stability assumptions (A3 and A4) in the estimation of ME^a by making weaker assumption (A10) (cf. Table 4.2).

4.3.4.3 Tests for measurement and response equivalence

Overall, the forward and backward methods in the WSCGD are either based on measurement (A5) or response equivalence (A6), besides the weaker assumptions (A8) to (A10) (Table 4.2). Our ability to test for equivalence across occasions based on a WSCGD is somewhat more limited, however, because, in case of non-equivalence, we only observe $E\left({}_1^bY_2^a\middle|{}_1^bS_2^a = 1\right)$ and $E({}_1^aY_2^a|{}_1^aS_2^a = 1)$ (i.e., the response mean at occasion two by modes assigned at occasion one), but not $E(Y_2^a|S_2^a = 1)$ (i.e., the response mean of respondents at occasion two who were not assigned to any mode at occasion one). In a WSCGD, it can be assessed, whether the assignment to modes A

and B at the first occasion, which is fully random, has an impact on the conditional means[14]

$$\Delta_4 := E\left({}_1^bY_2^a \middle| {}_1^bS_2^a = 1\right) - E({}_1^aY_2^a | {}_1^aS_2^a = 1). \tag{4.14}$$

If $\Delta_4 \neq 0$, (A5) and (A6) may not hold. This test, however, is not exact, because non-equivalence is tested as a compound. Furthermore, non-equivalence can only be detected, if its effects vary across modes assigned at the first occasion of the WSCGD. If

$$\begin{aligned} E\left({}_1^bY_2^a \middle| {}_1^bS_2^a = 1\right) &= E({}_1^aY_2^a | {}_1^aS_2^a = 1) \\ &\neq E(Y_2^a | S_2^a = 1) \end{aligned} \tag{4.15}$$

non-equivalence remains hidden. To avoid this problem it is possible to draw an additional independent sample surveyed only at the second occasion in mode A. Since this sample is fully independent from the first occasion, assessing

$$\Delta_5 := E(Y_2^a | S_2^a = 1) - E({}_1^aY_2^a | {}_1^aS_2^a = 1) \tag{4.16}$$

is superior to assessing significance of Δ_4, but requires additional effort in data collection[15].

4.4 Illustration

In 2011, Statistics Netherlands conducted a large-scale mixed-mode experiment (Buelens et al., 2012; Klausch et al., 2013a, 2013b; Schouten et al., 2013). Three separate WSCGDs were administered in parallel with a sample size of 6,803. Mode B was represented by telephone (n=1,659), mail (n=1,760), and web (n=1,746), respectively. Mode A was chosen to be face-to-face (F2F), where a sample of n=1,638 served as control. The survey topic was the national Crime Victimization Survey

[14] Since (A5) and (A6) are stated in terms of distributions, other moments may also be compared across conditions. We note that equivalence in expectations is most crucial for the estimators discussed in the present paper.

[15] Additional tests for equivalence can also be conducted based on exogenous information *X*. However, since *X* might often only be weakly related to survey variables, these tests are probably insufficient in practice.

(CVS). The intention was evaluating, whether a switch between F2F and telephone, mail or web would suggest a change in measurement error and if these modes would be compatible to a F2F follow-up (scenarios 1 and 2a).

4.4.1 Design considerations and response patterns

On average, 6 weeks lay between the first and second interview. From earlier implementations of the CVS, administered on yearly basis, it was known that many statistics only changed slowly or were stagnating, so that time-stability appeared plausible for this time frame. To reassure about this assumption, however, the control group was included.

Additionally, several precautions were taken in advance to assure equivalence of occasions. First, no reference to the second occasion was made at the first occasion including no mentioning of the possibility to reply later in a F2F mode. Interviewers, furthermore, were only vaguely informed about the second occasion while administering occasion one. This avoided that some individuals did not reply at the first occasion, because they preferred the forthcoming F2F mode at occasion 2. Second, the fieldwork at occasion two could hardly be distinguished from occasion one. No separate advance letter was sent at occasion two and instead interviewers contacted all individuals directly. F2F interviewers were instructed to recruit nonrespondents from occasion one as for a regular F2F survey and they were trained to explain respondents at occasion one the need for repeated participation for answering additional questions from the CVS. Finally, the standard CVS questionnaire was shortened for the re-interview to keep survey burden low. This measure may have challenged the measurement equivalence assumption. We return to this aspect in the discussion.

Figure 4.5 shows the missing data patterns of the three WSCGDs in analogy to the schematic illustration introduced in section 4.3.4. The upper part of the three plots is represented by the F2F control group (mode A; the same sample in the three designs, respectively). The lower parts show the nonresponse and response proportions of telephone, mail, and web (modes B, respectively). Clearly, mail (left) and telephone (right) achieved higher response (n=857 and n=735, respectively) than web (middle, n=497). Response was highest in F2F (n=1048).

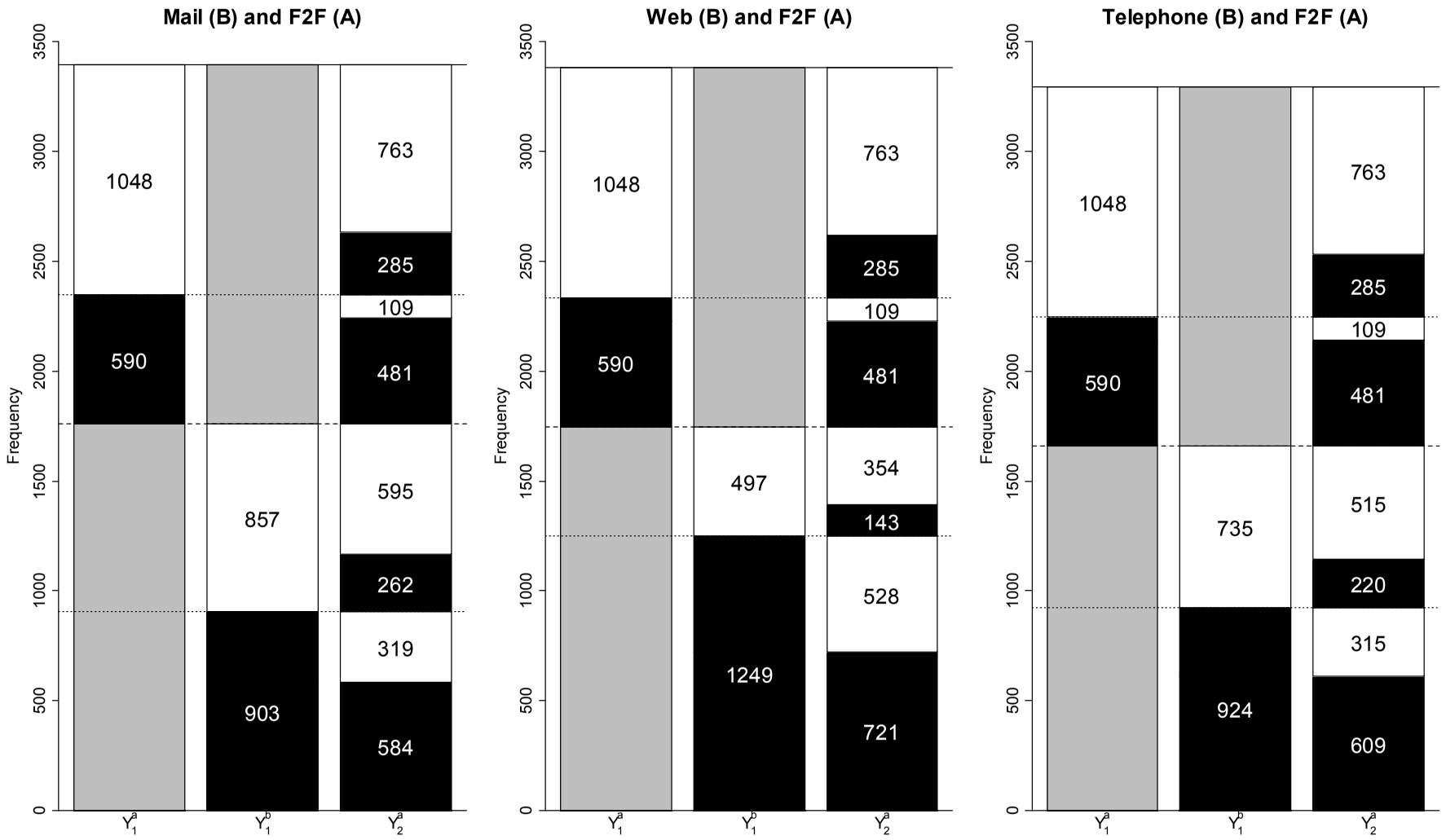


Figure 4.5: Missing data patterns of the three WSCG designs in the CVS experiment (face-to-face is mode A, respectively)

4.4.2 A comparison of two modes

We now illustrate estimation of (double-) conditional *MEs* and their selection biases (8) and (9) (selection effects) against ME_{BS}^{naive} (6), which could be estimated in between-subject designs. We start by comparing only two modes, web and F2F, and a single variable (middle missing data pattern in Figure 4.6). We selected a regularly reported CVS statistic called the 'Neighbourhood Decay Index'. It is based on multiple 3-point rating scale questions about deterioration of the neighbourhood aggregated and normalized to a summary score ranging from zero to ten.

We estimated the *MEs* and coefficients Δ using regression estimation (e.g., Imbens, 2004; Schafer & Kang, 2008). In addition to the target variables, eight socio-demographic indicators (*X*) were available[16]. The model was built using forward inclusion of those covariates that minimized the model AIC, where the target variable *Y* was always included in the first step[17]. This implies that for the forward method a linear model of $E\left(Y_2^{f2f}\middle|S_1^{web}=1,S_2^{f2f}=1,Y_1^{web},X\right)$ and for the backward method the model $E\left(Y_1^{web}\middle|S_1^{web}=1,S_2^{f2f}=1,Y_2^{f2f},X\right)$ was estimated. Based on the fitted models, outcomes $E\left(Y_2^{f2f}\middle|S_1^{web}=1,Y_1^{web},X\right)$ and $E\left(Y_1^{web}\middle|S_2^{f2f}=1,Y_2^{f2f},X\right)$ were predicted from observed data and the means $E\left(Y_2^{f2f}\middle|S_1^{web}=1\right)$ and $E\left(Y_1^{web}\middle|S_2^{f2f}=1\right)$ were estimated by averaging across the respective response sets. Standard errors were estimated using the bootstrap with 1,000 replications. At each replication a new adjustment model was fitted. The same procedure was applied in the control group[18].

We distinguish estimates from the WSD and the WSCD design (Table 4.3). Consider first the WSD, in which time-stability is assumed and cannot be tested or adjusted. We found significant (double-) conditional *MEs* suggesting that, both, web and F2F respondents provide more negative (i.e., higher) ratings in web than in F2F on the Decay Index. An unbiased estimate of ME_{BS}^{naive} is not available, because

[16] In particular: gender, age, income, civil status, nationality, household size, urbanity, and inhabitation of a large city in The Netherlands.

[17] Missing cases due to item nonresponse were list wise deleted in all analyses.

[18] This modelling approach is still relatively basic. In more advanced models, further variables Z could be included that support assumptions (A7a/b) and (A8a/b). For example, Schouten et al. (2013), using only the backward method, included survey attitudes and other related target variables from the second occasion in the model of $E(Y_1^b|S_1^b=1,S_2^a=1,Y_2^a,Z_2^a,X)$.

$E(Y_1^{f2f}|S_1^{f2f} = 1)$ cannot be estimated from a comparison mode. In the WSCGD, all effects including the adjustment coefficients Δ_1 to Δ_3 are available. The naïve *ME* was significant (.934) and comparable in magnitude to the *MEs* estimated from the WSD. However, we also found significant Δ_1 and Δ_2. A significant Δ_1 suggests that there was a change in answers for respondents in the control group across occasions (.265), which is used to adjust WSD estimate of ME^{web}. A significant Δ_2 suggests that it makes a difference[19], whether $E(Y_1^a|S_1^a = 1)$ (WSCGD) or $E(Y_2^a|S_2^a = 1)$ (WSD) is used in estimating ME^{f2f}.

Table 4.3: Overview on *ME* estimates from the within-subject and WSCG designs (mode A: F2F, mode B: web; Variable: 'Neighbourhood Decay')

	Within-Subject Design (WSD)			Within-Subject Control Group Design (WSCGD)		
	Est.	95% CI	p[a]	Est.	95% CI	p[a]
ME_{BS}^{naive}	-	-	-	.934	[.656,1.204]	<.001
ME^{F2F}	1.107	[.858,1.368]	<.001	.775	[.470,1.100]	<.001
ME^{web}	.966	[.731,1.192]	<.001	.701	[.418,.967]	<.001
$SE(Y_1^b)$	-	-	-	-.160	[.-381,.077]	n.s.
$SE(Y_1^a)$	-	-	-	-.234	[-.548,.100]	n.s.
Δ_1	-	-	-	.265	[-.406,-.129]	<.001
Δ_2	-	-	-	.385	[-.616,-.172]	<.001
Δ_3	-	-	-	-.053	[-.159,.051]	n.s.
Δ_4	-	-	-	-.062	[-.306,.169]	n.s.
$ME^{f2f,web}$	1.167	[.885,1.467]	<.001	0.875	[.885,1.467]	<.001
$ME^{web,f2f}$	.875	[.546,1.202]	<.001	1.167	[.546,1.202]	<.001

a: Bootstrapped p-value (1,000 draws) of a two-sided test against zero

After adjustment, we found that the bootstrapped difference between adjusted estimates from the WSCGD and WSD was significant (not shown in Table 4.3). The WSCGD estimates suggested somewhat smaller *MEs* and larger selection bias against the naïve

[19] This conjecture reflects that Δ_2 is a compound measure for the change of time Δ_1 and change of selection error on Y_2^a (cf. section 4.3.4). The significant Δ_1 to large extent implies significant Δ_2 in the present case.

estimator. The adjustment in the WSCGD thus proved to be important in this example. However, selection bias was not significant in this example.

Finally, an approximate test for equivalence (A5) and (A6) was available (Δ_4 , formula (4.15)). We found an insignificant Δ_4 suggesting that the type of mode offered at occasion one (web or F2F) did not have any impact on the conditional distribution of the Decay Index at occasion two. This was reassuring about measurement and response equivalence of occasions[20].

The conditional *ME* estimates suggest that respondents under web and F2F answered on the Decay Index with differing extents of measurement error in web and F2F. In the CVS, F2F was the standard mode of administration for a long time. In the hypothetical situation that the design would be switched completely to web (scenario 1), a break in this time series could be expected, because web respondents score more negatively on the index. If F2F is assumed as benchmark, furthermore, the web mode would evoke higher extents of error, both from respondents in web (ME^{web}) and F2F (ME^{f2f}). A sequential design combining web and F2F is not advisable, because the (double-) conditional *MEs* imply that respondents under both modes provide different answers. Due to the significant time-related change, the double-conditional *MEs*, which cannot be adjusted for change (cf. section 4.3.4), should be interpreted with care, however. Further research could evaluate, whether a unified mode design of the questions underlying the Decay Index makes it possible to reduce the (double-) conditional *MEs* of F2F against web.

4.4.3 Visualization of measurement effects for multi-mode comparisons

Another option to interpret the decomposition of naïve *MEs* into conditional *MEs* and selection bias (selection effects) is plotting all conditional means into interaction diagrams (Figure 4.6). This visualization has several advantages. Interaction plots give a quick overview on the size and type of conditional *MEs*, where parallel plotting allows a multi-mode comparison. Furthermore, the scale of all variables can be taken from the ordinate, whereas it cannot be taken from Table 4.3.

[20] Note that Δ_5 (16) cannot be calculated in this design, because an independent sample at the second occasion is not available.

Figure 4.6 shows the estimates for the Decay index across mail, web, and telephone from the WSCGDs. The middle plot visualizes the estimates from Table 4.3. Furthermore, the left plot illustrates the comparison of mail and F2F, and the right plot telephone and F2F. Each line represents a group of respondents in mode A or B. The estimates on the left side, furthermore, represent the conditional means of answers under mode A (Y_1^a, i.e., F2F), and the estimates on the right side the means of answers under mode B (Y_1^b, i.e. mail, web, or telephone). Now, the 'slope' of the two lines provides the conditional *MEs*, respectively, whereas the vertical distance between lines represents the selection bias. Selection bias and conditional *MEs* add up to the naive *ME* (formula (4.10)). All effects are indicated by labelled arrows.

It can be seen that, across modes, the conditional *MEs* differed. Whereas for web and mail we found negative *MEs* (significant) suggesting more negative rating of neighbourhood decay, the *MEs* of telephone against F2F were positive suggesting slightly better ratings of the neighbourhood than in F2F.

We illustrate analyses for two further key variables from the CVS (Figure 4.7 and Figure 4.8). The 'Social Quality Index' (Figure 4.7) is an index based on multiple questions about the social cohesion in the neighbourhood (normalized to a scale from 0 to 10). It can be seen that (significant) *MEs* of web and mail against F2F suggested more negative ratings of the neighbourhood, similar to the Decay Index. Telephone and F2F did not exhibit significant *MEs* for this variable. Another key variable in the CVS is the proportion of individuals having incurred a crime in the past twelve months ('victimization', Figure 4.8). Here, we found positive *MEs* in the mail and web conditions suggesting that a respondents in these modes reported more crimes on average than they would in F2F. However, the adjusted *ME* estimates did not reach significant levels for these cases[21].

[21] It should be noted that the number of repeated victimization questions in the follow-up was smaller than the set used for the standard statistics. The results may not generalize to the standard statistic for this reason. Schouten et al. (2013) present estimates for victimization in the past 12 months, similar to ours. Their estimates are based on a slightly larger sample size and the full set of victimization questions. The estimates for ME^{F2F} are -3.3 ($p<.05$), 3.5 (n.s.), and 3.6 (n.s.) for telephone, mail, and web, respectively, which suggest the same tendencies as the estimates presented here based on the reduced set and a slightly smaller sample.

In summary, these graphs show a situation, in which the self-administered modes exhibit stronger *MEs* against F2F than the interviewer-administered telephone mode, possibly owing to the role of the interviewer in the response process (De Leeuw, Dillman, et al., 2008; De Leeuw, 1992). In scenarios 1 and 2a (with F2F benchmark), sequential mixed-mode designs and single-to-single mode switches involving only F2F and telephone modes, therefore, seem more realistic for the given sets of questions, but the F2F mode should not be combined or exchanged with a self-administered mode, because these provide larger measurement error compared to the F2F benchmark. Unified mode design of questions underlying the three indices would be necessary to reduce the observed *MEs* before implementing such redesign strategies.

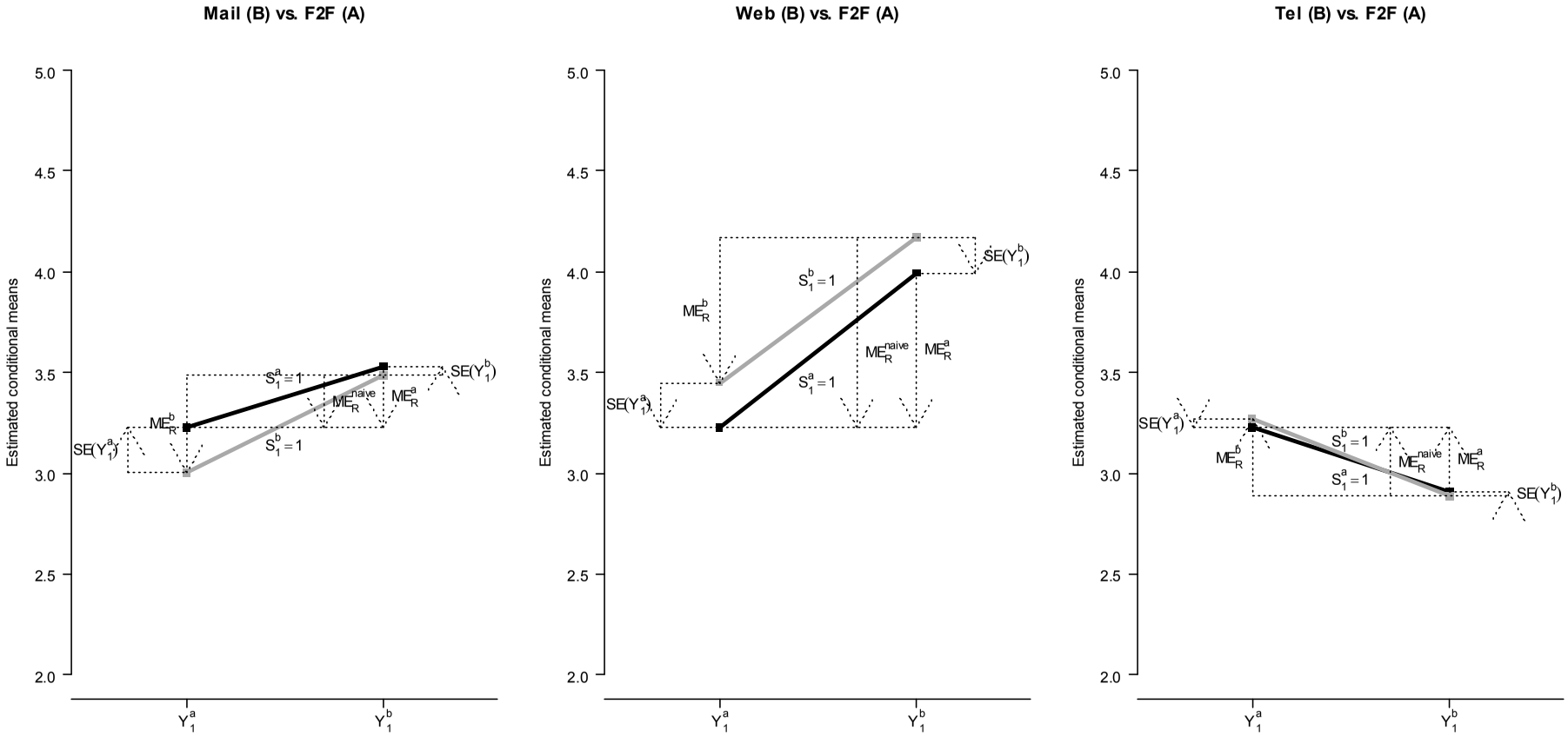


Figure 4.6: Interaction diagrams of conditional *MEs* (significant) and their selection bias against the naïve *ME* estimator (not significant) for the Neighbourhood Decay Index of three modes (Mail, web, and telephone) against face-to-face (mode A). All (adjusted) estimates based on the WSCGD (F2F respondents: black; mail, web, and telephone respondents: grey)

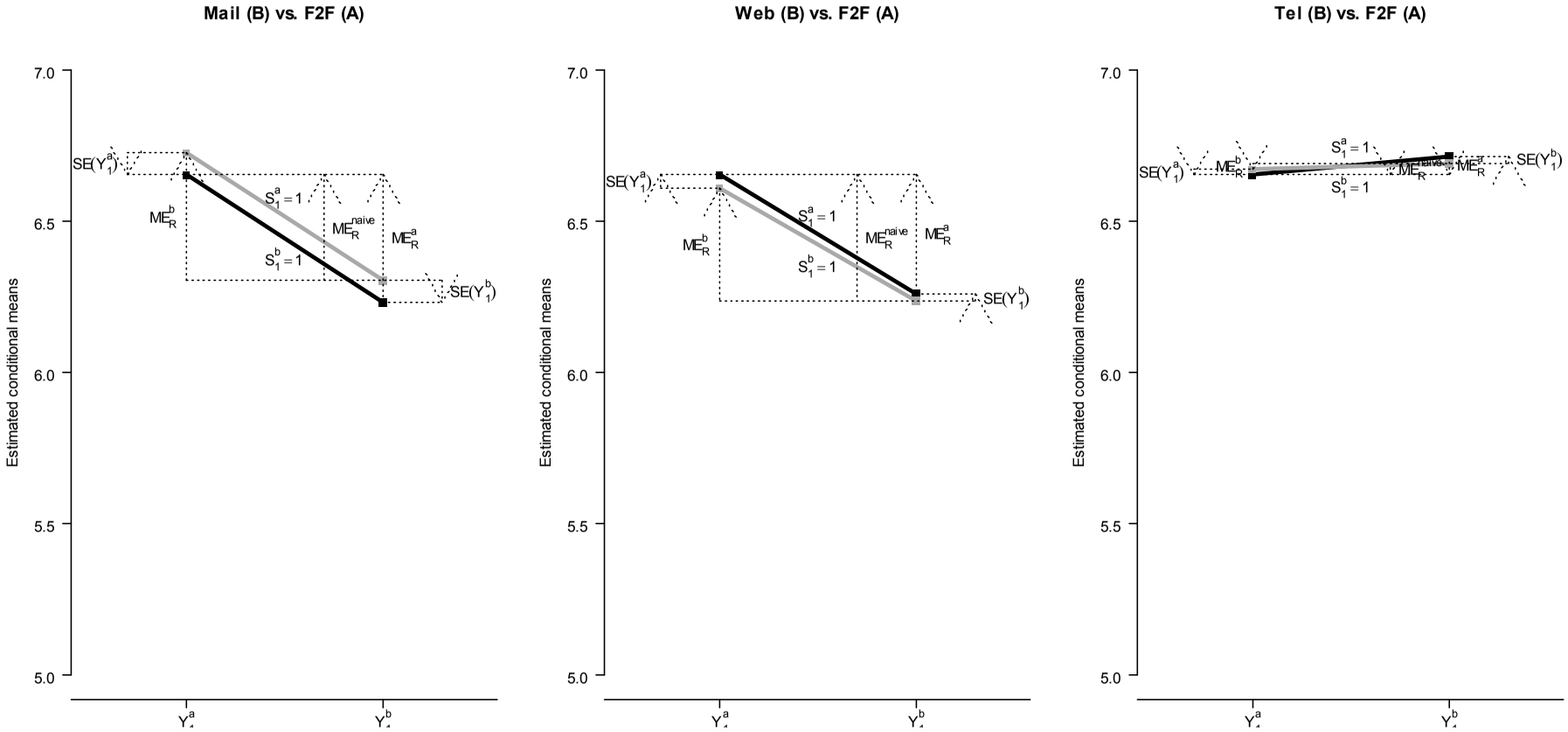


Figure 4.7: Interaction diagrams for the Social-Quality Index based on the CVS experiment (The *MEs* of mail and web reached significant level, whereas telephone showed insignificant *MEs*). All (adjusted) estimates based on the WSCGD (F2F respondents: black; mail, web, and telephone respondents: grey)

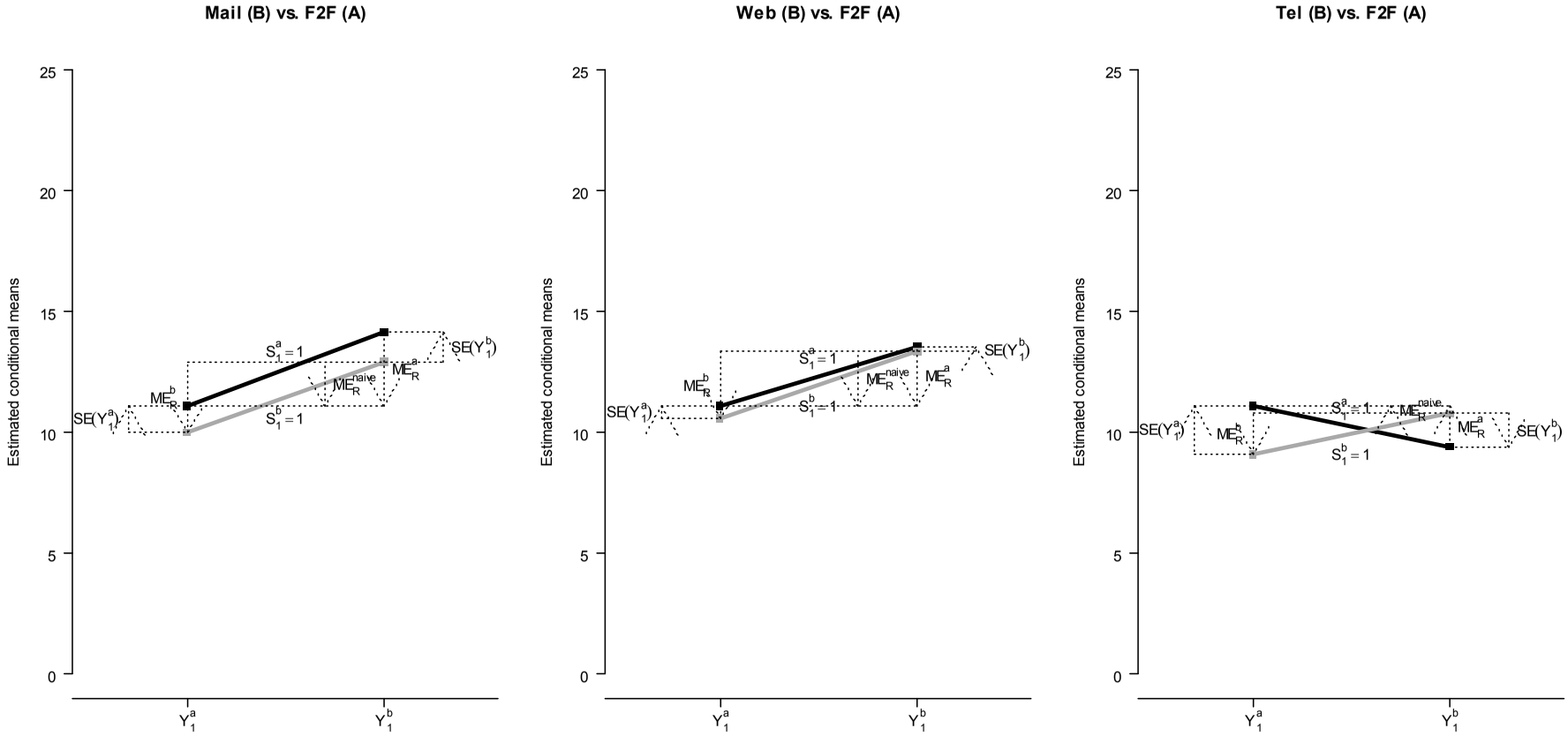


Figure 4.8: Interaction diagrams for the proportion of Victimization (in %) based on the CVS experiment (*MEs* are insignificant in all cases). All (adjusted) estimates based on the WSCGD (F2F respondents: black; mail, web, and telephone respondents: grey)

4.5 Discussion

A new framework was introduced which systematizes required assumptions for using within-subject designs (WSDs) to estimate *MEs* relevant in two mixed-mode redesign scenarios: single-to-single mode switches and introduction of sequential mixed-mode surveys. A key advantage are weaker MAR assumptions than earlier approaches commonly make, when estimating *MEs* in between-subject or sequential mixed-mode designs based on comparatively weak auxiliary information. Although we did not identify significant selection bias in the illustration, the additional auxiliary information in WSDs re-assures that estimated *MEs* are causal to naïve *ME* estimates and selection bias is excluded as an alternative explanation.

The conditional and double-conditional *MEs* needed for judging unified mode designs in scenarios 1 and 2 are defined by Table 4.1. These *MEs* denote differences in measurement error conditional on different sub-sets of respondents. Therefore, the quantities may be unequal and lead to different conclusions about the size of measurement error differences. In interaction diagrams this is indicated by diverging or converging lines. Conclusions about the scenarios may then depend on the type of *ME* that is required. We found an example of an interaction effect on a question about the safety perception of the neighbourhood (Appendix C), where ME^a was significantly larger than ME^b in mail and web. However, other interactions were absent in the CVS data.

A distinction has to be made between assumptions in WSDs and WSCGDs (Table 4.2). WSDs strongly depend on time-stability assumptions, which can be relaxed and adjusted in WSCGDS. In the WSCGD, measurement and response equivalence still is required. The backward method estimates ME^a (and $ME^{a,b}$) assuming response equivalence and the forward method estimates ME^b (and $ME^{b,a}$) assuming measurement equivalence. Equivalence can only be tested approximately and thus mainly depends on plausibility arguments. It should be noted that when the order of modes is exchanged in the WSCGD, thus, contrary to the order introduced in section 4.3, A (e.g., F2F) is administered before B (e.g., web), estimating ME^b would assume response equivalence and ME^a measurement equivalence reversing the two assumptions. By sequencing modes, the type of equivalence assumption that is desired

and deemed more plausible can therefore be chosen. This raises the important question, which WSCGDs are feasible for which modes in terms of backward and forward estimation and related equivalence assumptions. This issue is mainly open to further research, but we note the following aspects.

First of all, it appears natural to follow the order of the sequential mixed-mode design, when *ME* estimates are used for designing a sequential mixed-mode survey (scenario 2). When scenario 1 is of interest, however, the order 'B to A' or 'A to B' appears artificial at first, but needs to be based on the following practical considerations.

In the forward method, repetition of target variables similar to test-retest designs is required to assure measurement equivalence, but response equivalence is not central. Longer time-lags appear particularly important to allow answers provided at an earlier occasion to be forgotten and more difficult to be reproduced. Future research could evaluate which time lags make measurement equivalence more likely. Additionally, some reduction of response burden may also be required (e.g., by shortening questionnaires), because a strong reduction in follow-up response may cause estimation problems (e.g., small samples). An important question is consequently, to which degree measurement equivalence can still be assured when questionnaires are shortened.

An advantage of the backward method in WSCGDs is that an exact repetition of survey questions is not necessary as long as response equivalence can be assured. Any type of variable can be collected in the re-interview to turn the backward MAR assumption plausible, because variables from occasion two are only used in nonresponse adjustment but not for estimating means. Using the data from the CVS experiment, Schouten et al. (2013) estimated ME^{f2f} (ME^{a}) by calibration to re-interview F2F response. This procedure can be classified as backward method in the WSCGD framework. The authors used a shortened version of the CVS to keep response burden low and increase chances for response equivalence. Additionally, they surveyed other (non-CVS) variables that were used in nonresponse adjustment. However, in the scenarios of interest in the illustration (1 and 2a with F2F as benchmark), ME^{b} (e.g., ME^{web}) is actually the relevant quantity, which, in the CVS experiment, requires estimation by the forward method. If ME^{f2f} (ME^{a}) is estimated (e.g., because measurement equivalence cannot be assumed) but ME^{b} is of interest, it has to be

assumed that both quantities are approximately equal and interaction effects thus are negligible. Future research could investigate how frequently such interactions are absent in practice.

An alternative may be to reverse order of modes in replications of the experiment (e.g., F2F followed by web, mail, or telephone). Since this procedure reverses assumptions, ME^b could be estimated by the backward method (e.g., ME^{web}). However, it would need to be evaluated, whether response equivalence would be given in the follow-up modes. It has to be considered that interviewer modes (e.g., F2F) may outperform self-administered modes as follow-up modes, because interviewers can impact stronger persuasive power in recruiting respondents for a re-interview.

Finally, in designing WSCGDs we note a trade-off between costs, sample size, and related power. Clearly, a WSD yields approximately equivalent costs as a between-subject design. However, the control group can strongly increase costs, especially, if based on an expensive mode like F2F. In the illustration, a sample of approximately 1,700 units per condition led to rather large standard errors (Table 4.3). It is therefore important to assess required *n* for acceptable detection power of *MEs* vis-à-vis acceptable costs. Standard errors of estimates are also influenced by the efficiency of the estimation method. Currently, there is ongoing debate beyond this discussion, on the 'best' estimation method for parameters and their standard errors when observations are missing at random.

WSDs may prove to have great potential for mixed-mode survey design. In the future, re-interview data may be used in many situations including sequential mixed-mode surveys and mixed-mode panel surveys to estimate and adjust *MEs*. An adjustment method could predict or impute plausible values for potential outcomes not observed in a mixed-mode design, for example. As long as it is possible to collect repeated measurements a correct model for *MEs* can form the basis of adjustments. The suggested framework may guide researchers in developing and extending this methodology.

5 Evaluating Bias of Sequential Mixed-Mode Designs against Benchmark Surveys [1]

Sequential mixed-mode designs have become an important alternative to single-mode designs in international survey research. Sequential designs provide nonrespondents or non-covered persons in a single-mode survey (e.g., web) with another opportunity to respond by offering at least one other mode as a response option (e.g., face-to-face). This follow-up can increase response and coverage while achieving cost efficiency by sequencing inexpensive before expensive modes (De Leeuw, 2005; Dillman, Smyth, et al., 2009; Lynn, 2013).

Since there is a direct trade-off of mode-specific errors in the total survey error (TSE) of sequential mixed-mode designs, mixed-mode surveys may offer an optimal error balance (Bethlehem & Biffignandi, 2011, p. 235; De Leeuw, 2005; Groves et al., 2010, p. 176). Selection errors (i.e., coverage error or nonresponse error) and measurement errors have received particular attention in this context, because they are often systematic (i.e., they cause bias in linear estimates, such as means) and differ in size across modes (Biemer & Lyberg, 2003, p. 59; De Leeuw, Dillman, et al., 2008; De Leeuw, 2005, 2008; Klausch et al., 2013b; Kreuter et al., 2008). In practice, survey designers need to estimate these biases to monitor survey accuracy and to decide about the impact of changes to fieldwork or questionnaire designs on accuracy. In doing so, two basic research questions (RQs) are of concern:

[1] This chapter is under review as Klausch, T, Hox, J., Schouten, B., Evaluating Bias of Sequential Mixed-Mode Designs against Benchmark Surveys.

First (RQ1), is the mixed-mode survey needed or would a single-mode survey suffice in terms of accuracy? Given equal budget and time constraints the design featuring lowest survey error is preferred (Biemer & Lyberg, 2003, p. 44; Biemer, 2010b; Groves & Lyberg, 2010; Kreuter et al., 2010). To evaluate the accuracy of a mixed-mode design, the expected total bias needs to be estimated for different candidate mixed-mode designs and compared to the error of single-mode surveys without the mixed-mode follow-up. The mixed-mode survey is only needed, if the total bias of the single-mode survey is decreased by the follow-up.

Second (RQ2), what are the major systematic sources of error in single-mode surveys and how are they impacted by the mixed-mode follow-up? The primary motivation for sequential mixed-mode surveys is the reduction of selection bias. If the size of selection bias before and after the follow-up is estimated, the actual need and success of reducing selection bias can be evaluated. However, if one mode in a mixed-mode design measures with less accuracy, reductions in selection bias may be offset by increases in measurement bias of the mixed-mode estimate. To prevent this problem, questionnaires of the mode evoking higher measurement bias can be redesigned once the size of bias is known (Dillman, Smyth, et al., 2009, p. 326).

In this paper, we answer these RQs for the case of the Dutch Crime Victimization Survey, a national survey conducted by Statistics Netherlands, based on a large-scale experiment with three mixed-mode designs: web, mail, and telephone followed up by face-to-face, re-analysing data collected by (Schouten et al., 2013). We estimate total bias of the modes before and after the mixed-mode follow-up (RQ 1) and decompose it into selection and measurement bias components (RQ 2). A major problem in the evaluation of survey bias is its estimation. Past research has discussed a series of approaches to this problem. The next section reviews this research and outlines the approach of the present study.

5.1 Background

Past research has mainly discussed three different approaches to evaluating bias in mixed-mode surveys: the record-check approach, the relative mode-effect approach, and the benchmark survey approach. Exact bias estimation is only possible using the record-

check approach, which supposes that true scores are available from an external database (Biemer & Lyberg, 2003, p. 289; Groves, 2006). A growing number of case studies evaluate mode differences in bias using record checks (Kirchner & Felderer, 2013; Körmendi, 1988; Kreuter et al., 2013, 2008; Olson, 2006; Sakshaug et al., 2010; Tourangeau, Groves, & Redline, 2010) and the approach is also applied to mixed-mode surveys (Fowler et al., 2002; Kreuter et al., 2010; Link & Mokdad, 2006; Sakshaug et al., 2010; Voogt & Saris, 2005). In practice, however, records are hardly ever available for the survey variables of interest and involve various other practical problems, such as incomplete databases or matches to units[2] (Biemer & Lyberg, 2003, p. 291; Körmendi, 1988; P. V. Miller & Groves, 1985).

Vannieuwenhuyze and Loosveldt (2013) suggested evaluating mixed-mode surveys by 'relative mode effects' (Vannieuwenhuyze et al., 2010, 2014). Relative effects are defined as difference in total bias, measurement bias, and selection bias between mode-specific response groups in a mixed-mode design, referred to as overall, measurement, and selection effects. However, relative mode effects cannot be interpreted in an absolute sense. For example, a relative selection effect in a sequential mixed-mode design does not suggest that selection bias is reduced or increased by the follow-up, but only that different 'types' of respondents are reached.

This problem is addressed when biases are evaluated against a 'preferred' single-mode survey whose measurements are considered valid and which is also considered optimal on selection bias (Biemer & Lyberg, 2003, p. 287; Biemer, 1988; De Leeuw, Dillman, et al., 2008; De Leeuw, 2005; Körmendi, 1988; Vannieuwenhuyze et al., 2010; Vannieuwenhuyze, 2014). This mode is called the 'single-mode benchmark' (SMB). A relevant application of SMBs is the redesign of repeated cross-sectional surveys to mixed-mode designs. To assure comparability, the change in bias relative to established SMB time-series is an important concern (addressing RQ 1). Furthermore, preventing

[2] Records may not coincide with the study time period and may themselves contain errors, survey questions may suffer from specification problems compared to information encoded in records, and there may be problems arising from incomplete matches of respondents to records (Biemer & Lyberg, 2003, p. 291; P. V. Miller & Groves, 1985). Furthermore, records are seldom available to all researchers due to privacy limitations and sometimes can concern very specific sub-populations, such as students (Kreuter, Presser, & Tourangeau, 2008).

shifts after assessing selection and measurement effects against the SMB may be possible (RQ 2).

Using the SMB approach, Schouten et al. (2013) and Klausch, Schouten, and Hox (2014) defined differences in bias between a SMB and other modes as 'single-mode effects'. The primary difference to the relative mode effect approach is that effects are defined between single-mode surveys and a reference survey (the SMB), and not between mode-specific response groups in a mixed-mode design. In an empirical study of the Crime Victimization Survey, the authors used a face-to-face survey as SMB against which telephone, mail, and web modes were compared. However, in evaluating sequential mixed-mode surveys (RQ 1 and 2) against a SMB, the impact of the follow-up mode on single-mode effects needs to be considered as well, which has not been accomplished so far. The bias of a mixed-mode survey against the benchmark is called a 'mixed-mode effect'. This extension is provided in the present study.

A limitation of the SMB approach is that single surveys may not be fully appropriate as benchmark. For the case of face-to-face, for example, we assume that due to generally high response rates selection bias is acceptable, but face-to-face measurements may be considered too erroneous to be used as benchmark, for example, due to social desirability bias (Tourangeau et al., 2000, p. 257; Tourangeau & Yan, 2007). In particular, self-administered modes, such as 'web', may be more precise for sensitive questions (Kreuter et al., 2008). In this case, a combined benchmark of the selection bias of face-to-face and measurements of web may appear superior to the SMB. We call this combined benchmark the 'hybrid-mode benchmark' (HMB). Even though the SMB has a long-standing tradition in evaluating survey errors (Biemer & Lyberg, 2003, p. 291), the HMB may be more immediate to survey practice when assuring comparability to time series is not central (e.g., when questions are sensitive or interviewer effects are strong in the SMB). The present paper is the first to consider a HMB besides the SMB.

Previous literature also established the general conditions under which unbiased estimation of measurement and selection effects is possible (Klausch et al., 2014; Vannieuwenhuyze et al., 2014; Vannieuwenhuyze & Loosveldt, 2013). The primary difficulty in estimation is the confounding of both effects in the overall mode effect. To

disentangle the effects, exogenous auxiliary information needs to be available, conditional on which measurements and the selection mechanism into mode-specific response groups are independent (Imbens, 2004; Morgan & Winship, 2007; Pearl, 2009; Rubin, 2005). In mixed-mode surveys, finding variables that allow both unconfoundedness and exogeneity appears difficult, however. To address this problem, Schouten et al. (2013) and Klausch et al. (2014) introduced a re-interview design. In the re-interview, repeated measures of survey target variables as well as other auxiliary information are collected. Conditional on the repeated measures, the unconfoundedness assumption appears more plausible (cf. section 5.3). The present study re-analyses this data for the case of sequential mixed-mode surveys using both a SMB and HMB.

5.2 The Crime Victimization Survey Case Study

The case study was conducted in the context of the Crime Victimization Survey (CVS), administered by Statistics Netherlands in 2011 in an experiment conducted independently from the regular CVS (Klausch et al., 2013b, 2014; Schouten et al., 2013). A face-to-face survey (F2F in the following) served as SMB and was administered in parallel to three different single-mode surveys (telephone, mail, and web). The F2F survey showed the largest response rates of the four modes (64.5%; Table 5.1). In addition, Klausch, Hox, and Schouten (2013a) showed that representativeness of F2F on socio-demographic background variables was high in the CVS experiment. It can, therefore, be considered a suitable benchmark for selection bias. Furthermore, predecessors of the CVS were F2F and the measurements from F2F are by many regarded accurate or desirable. Setting F2F as SMB appears plausible from this point of view.

In addition to the single-mode designs, the experiment entailed a mixed-mode component re-approaching the nonrespondents in telephone, mail, and web by F2F. This procedure may yield mixed-mode estimates that are similar to the F2F benchmark and provide inexpensive designs compared to F2F alone. Evaluating the success of this procedure is the objective of the present study (RQs 1 and 2).

Table 5.1: Response rates and mixture weights in the CVS mixed-mode experiment (weighted)

	F2F (SMB) % (n=1,639)	Telephone[a] % (n=1,658)	Mail % (n=1,760)	Web % (n=1,746)
Single-mode	64.5	48.6	49.3	29.0
Mixed-Mode	-	65.2	67.3	59.7
Prop. Single-mode Resp. (π)	-	74.6	73.3	48.6
Prop. Follow-up Resp. ($1-\pi$)	-	25.4	26.7	51.4
Re-interview (full 2nd wave)	53.9	50.6	52.4	51.4

a. The telephone response rates are taken against the net sample of all sampled units (including persons without known telephone number).

In addition, we provide alternative analyses for the HMB case, discussing differences in conclusions about RQs 1 and 2 for both cases. The SMB assumption that the F2F mode could also serve as a benchmark for measurements can be questioned. The CVS contained a large number of attitudinal and sensitive questions, which may be particularly susceptible to stronger measurement bias due to the presence of an interviewer (cf. section 5.1). Measurement in the anonymous situation of self-administration (e.g., web) may be more appropriate for many CVS questions (cf. Appendix D for an overview on all questions). Nevertheless, F2F appears as a valid selection benchmark. The HMB thus assumes F2F as the benchmark for selection and web as the benchmark for measurement.

Each survey was based on a probability sample drawn from the national register[3] (a person sampling frame). The sample size was chosen such that minimal observable total single-mode bias against F2F is equal to the required precision of the CVS at the national level. The regular CVS is much larger because of detailed publications for a range of subpopulations.

[3] The original sample size was 2,200 persons in each mode-specific condition. However, only approx. 80% of the sample was followed up by F2F for cost-related reasons. In this sub-sample, all persons without registered telephone number were followed up, which lead to a slight over-representation of non-telephone households in the sub-sample. The response rates and subsequent analyses use design weights to yield unbiased estimates. In general, the deviations between weighted and unweighted results were small however. Furthermore, 500 respondents from the first wave of interviews were classified as not re-approachable in the second wave during fieldwork of the first wave (e.g., due to long-term illness or hard refusal). A stratified sample (by mode and telephone coverage) of 399 respondents was assigned to the sample of the second wave and treated as wave 2 nonrespondents. In chapters 2 of this dissertation the same procedure was applied, but a different random sub-sample was used. Therefore, the response rates in Table 5.1 marginally deviate from the response rates reported in Table 2.1.

The response rates in telephone, mail, and web were lower than in F2F (Table 5.1), but the F2F follow-up increased the response rates in all modes (up to 65.2, 67.3, and 59.7%, respectively). Since the single-mode response rate in the web mode was lower than in telephone or mail conditions, the relative proportion of F2F respondents in the web mixed-mode sample was substantially higher (51.4% vs. 25.4% and 26.7%, fourth row Table 5.1). These relative proportions (called π) suggest that the estimates from the mixed-mode web sample may be impacted more strongly by the follow-up than mixed-mode mail or telephone. It is, therefore, relevant in estimation, discussed in the next section.

5.3 Definition and Estimation of Single- and Mixed-Mode Effects

In this section, we explain how single- and mixed-mode bias components required to answer the RQs were estimated using the single-mode (SMB) and hybrid-mode (HMB) benchmark. We first define all biases following the total survey error (TSE) framework, but in a mixed-mode context some extensions of the TSE notations are necessary (Biemer & Lyberg, 2003; Biemer, 2010b; Groves & Lyberg, 2010; Groves et al., 2010, p. 48). Subsequently, we explain estimation of the biases against the SMB and HMB.

5.3.1 Defining single- and mixed-mode bias of the sample mean

To simplify notation, we define the bias of a mixed-mode telephone - F2F survey, but the definitions for the other modes follow likewise. We are interested in the bias of the estimator of the response sample mean of a continuous or discrete survey variable Y denoted by

$$\mu^{tel} = E(Y^{tel}|S^{tel} = 1),$$

where Y^{tel} represents the measurement of Y by telephone. The binary random variable S^{tel} represents the response mechanism of telephone and $S^{tel} = 1$ indicates the group of respondents. Let $\hat{\mu}^{tel}$ be the estimator of the sample mean. Its total bias (TB) against the population mean $\mu = E(Y)$ is (addressing RQ 1)

$$B(\hat{\mu}^{tel}) = TB^{tel} = \mu^{tel} - \mu. \quad (5.1)$$

Of course, μ is not observed without external validation data. When using the SMB and HMB approach, the population mean, therefore, will be substituted by an estimator that is least biased based on the respective benchmark assumptions (section 5.3.2).

The TB can now be decomposed into its systematic components, where we distinguish selection (SB) and measurement bias (MB). In limiting our elaboration to these biases, we also assume that other sources of bias discussed under the TSE framework, such as specification and data processing error, are negligible or at least equal across modes. Then measurement and selection bias can be said to add up to 'total bias'. Single-mode SB and MB follow as (addressing RQ 2)

$$SB^{tel} = E(Y|S^{tel} = 1) - \mu \quad (5.2)$$

and

$$MB^{tel} = E(Y^{tel}|S^{tel} = 1) - E(Y|S^{tel} = 1). \quad (5.3)$$

It can be seen that, following the TSE framework, SB is defined as the difference in sample mean of the true score and the population mean, whereas MB represents the difference of sample mean of the true score and measured mean answers[4]. MB and SB add up to TB.

In the mixed-mode survey, the answers of follow-up respondents in F2F are added to the single-mode response set. The mean from the pooled mixed-mode survey can then be regarded as the mean of a mixture distribution of Y^{tel} and Y^{F2F},

$$\mu^{MM} = \pi\mu^{tel} + (1 - \pi)E(Y^{f2f}\big|S^{tel} = 0, S^{f2f} = 1), \quad (5.4)$$

where the mixture constant π is defined by the expected proportion of single-mode respondents introduced in section 5.2 (cf. Table 5.1 for sample estimates from the case

[4] It should be noted that selection bias treats nonresponse and coverage bias as a compound (Vannieuwenhuyze & Loosveldt, 2013), which allows studying errors of modes by two fundamentally different processes (i.e., measurement and selection). Schouten et al. (2013) discuss estimation of single-mode coverage and nonresponse bias against a F2F benchmark.

study). The total bias of the mixed-mode mean now can be expressed as (addressing RQ 1)

$$B\left(\hat{\mu}^{MM}\right) = TB^{MM} = \pi TB^{tel} + (1-\pi)TB^{FU}, \tag{5.5}$$

where

$$TB^{FU} = E(Y^{f2f}\big|S^{tel} = 0, S^{f2f} = 1) - \mu \tag{5.6}$$

represents the total bias of the follow-up mode. It can be seen that the relative impact of single-mode biases is reduced by the size of the response proportion π, but the sign and size of bias of the follow-up sample is crucial for the overall mixed-mode TB (for a numerical example see Bethlehem & Biffignandi, 2011, p. 259). The mixed-mode SB and MB depend on the (weighted) follow-up mode bias in the same manner, i.e. (addressing RQ 2)

$$SB^{MM} = \pi SB^{tel} + (1-\pi)SB^{FU}, \tag{5.7}$$

where

$$SB^{FU} = E(Y\big|S^{tel} = 0, S^{f2f} = 1) - \mu, \tag{5.8}$$

and

$$MB^{MM} = \pi MB^{tel} + (1-\pi)MB^{FU}, \tag{5.9}$$

where

$$MB^{FU} = E(Y^{f2f}\big|S^{tel} = 0, S^{f2f} = 1) - E(Y\big|S^{tel} = 0, S^{f2f} = 1). \tag{5.10}$$

In the next section, we discuss how these biases can be estimated. We address estimation against a SMB and HMB in turn.

5.3.2 Estimation of single- and mixed-mode effects against the SMB

Setting a benchmark implies choosing two components. First, a mode is chosen that substitutes the true scores of *Y* by the measurements of the benchmark mode. These

observed scores are close or equal to the true scores (cf. section 5.1). Second, a mode is chosen that evokes acceptable selection bias. While acknowledging that the selection benchmark may not be free of selection bias (because itself suffers from unit nonresponse), it is considered the mode with best selection properties for the variable at hand. In principle, the measurement and selection benchmark modes may differ, in which case the combined benchmark is called hybrid. The SMB is the special case when both components are taken from the same mode.

In the present study, the F2F survey represents the SMB. In doing so, we set the F2F answers as the true score to assess comparability of single- and mixed-mode estimates of telephone (or the other modes, RQ 1) as well as reasons for incomparability (RQ 2). Thus, the population parameter μ is substituted by the F2F mean $E(Y^{f2f}|S^{f2f}=1)$. From the split-ballot mixed-mode design of the case study, the single- and mixed-mode total biases can now be estimated as (addressing RQ 1):

$$T\hat{B}_{SMB}^{SM} = \hat{E}(Y^{tel}|S^{tel}=1) - \hat{E}\left(Y_1^{f2f}\middle|S_1^{f2f}=1\right), \tag{5.11}$$

and

$$T\hat{B}_{SMB}^{MM} = \pi T\hat{B}_{SMB}^{SM} \tag{5.12}$$
$$+(1-\pi)\left(\hat{E}\left(Y_2^{f2f}\middle|S^{tel}=0, S_2^{f2f}=1\right) - \hat{E}\left(Y_1^{f2f}\middle|S_1^{f2f}=1\right)\right),$$

where $\hat{E}$ denotes the response means of F2F, telephone, and the follow-up. Furthermore, the indices '1' and '2' indicate that measurements or selection mechanisms from the single-mode F2F sample or the follow-up sample are used, respectively. The estimates can also be called 'total mode effects' or, alternatively, 'overall mode effects' or 'mode system effects' (Biemer, 1988; De Leeuw, 2005; Schouten et al., 2013; Vannieuwenhuyze & Loosveldt, 2013). However, in answering RQ 1 it is important to distinguish between the single- (5.11) and the mixed-mode effect (5.12).

The primary difficulty in estimating the single-mode selection and measurement effect components of the total effect (addressing RQ 2) is that after substitution of Y_1^{f2f} in (2) and (3) the outcome

$$E\left(Y_1^{f2f} \middle| S^{tel} = 1\right)$$

is not observed in a simple comparative mixed-mode design. This can be seen when illustrating the missing data pattern for the SMB and mixed-mode samples (Figure 5.1, left hand).

In the SMB sample, respondents provide answers on Y_1^{f2f} (white area) and unit nonresponse (black area). Also in the mixed-mode sample, the first part of the mixed-mode survey leads to telephone answers Y^{tel} and nonresponse, but additionally, the F2F follow-up of nonrespondents in telephone leads to responses Y_2^{f2f}. The grey areas indicate the part of missing information that is not observed by design. These outcomes can be called 'potential' following terminology introduced by Rubin (Klausch et al., 2014; Rubin, 1978, 2005; Vannieuwenhuyze & Loosveldt, 2013).

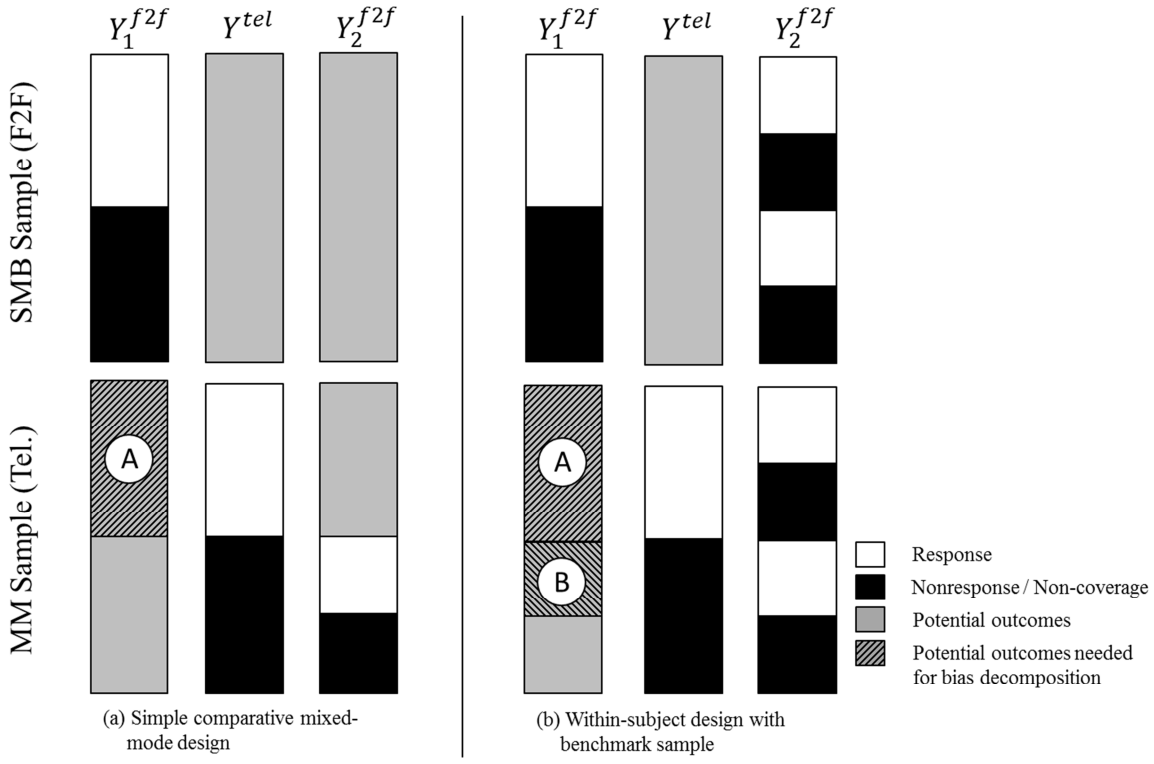


Figure 5.1: Missing data pattern of a simple comparative mixed-mode design (a) and a within-subject design (b) for estimating bias components against a SMB (Example of a telephone-F2F mixed-mode design with F2F benchmark)

$E\left(Y_1^{f2f} \middle| S^{tel} = 1\right)$ represents the mean of a subset of potential outcomes (i.e., telephone respondents), indicated by shaded area A. In the simple comparative design (left hand, Figure 5.1), additional auxiliary information is needed to estimate this mean (Vannieuwenhuyze et al., 2010, 2013). In the case study, we, therefore, applied a re-interview design called within-subject design (right hand; Schouten et al., 2013; Klausch et al., 2014; cf. re-interview response rates in Table 5.1). In the design, questions from the first occasion were repeated in the mixed-mode and the SMB samples. Using this data, we estimated $E\left(Y_1^{f2f} \middle| S^{tel} = 1\right)$ by multiply imputing potential outcomes Y_1^{f2f} (Little & Rubin, 2002; Rubin, 1987; Schafer & Graham, 2002). Peytchev (2012) discusses advantages of imputation for estimation of nonresponse and measurement bias in similar survey designs.

In doing so, we assume that potential outcomes and unit nonresponse at the re-interview are missing at random[5] (MAR) given observed outcomes Y^{tel} and Y_2^{f2f}. The repeated measures play an important role in making this assumption plausible (Schouten et al., 2013). The telephone measurements Y^{tel} are replications of Y_1^{f2f} given a measurement effect, but the partial correlation of both variables is unknown (cf. Figure 5.1, right side). However, from the re-interview data, Y_2^{f2f}, partial correlations to both Y^{tel} and Y_1^{f2f} can be estimated[6]. In addition, Y_2^{f2f} are closely related to these variables, which is an important criterion for the adequacy of auxiliary variables in causal inference and nonresponse adjustment (Bethlehem et al., 2011, p. 249; Schafer & Kang, 2008). The details of the practical implementation of the imputation procedure are provided in section 5.3.4.

---

[5] For the benchmark mode, we assume

$P\left(Y_1^{f2f} \middle| {}^{obs}Y_1^{tel}, {}^{obs}Y_2^{f2f}, S_1^{f2f} = 1\right) = P\left(Y_1^{f2f} \middle| {}^{obs}Y_1^{tel}, {}^{obs}Y_2^{f2f}, S^{tel} = 1\right) =$
$P\left(Y_1^{f2f} \middle| {}^{obs}Y_1^{tel}, {}^{obs}Y_2^{f2f}, S^{tel} = 0, S_2^{f2f} = 1\right)$. This assumption is equivalent to requiring the response mechanism of F2F response relative to mixed-mode telephone-F2F response to be conditionally unconfounded with Y_1^{f2f} (Imbens, 2004).

[6] If only Y_1^{tel} was available for imputing potential outcomes, unbiased estimation was only possible, if the response mechanism to F2F or mixed-mode telephone was strongly ignorable (Y_1^{f2f} was missing completely at random), because Y_1^{f2f} and Y_1^{tel} do not overlap (i.e. their partial correlation is unknown).

After imputation, the potential outcomes $E\left(Y_1^{f2f}\middle|S^{tel}=1\right)$ were estimated by the mean of imputed F2F outcomes for the telephone response sample (formulas (5.2) and (5.3)). Mixed-mode selection and measurement effects against the SMB were estimated in an analogous way. This task required estimating the mean of benchmark answers of follow-up respondents (formulas (5.8)and (5.10)) using the second potential outcome

$$E\left(Y_1^{f2f}\middle|S^{tel}=0,S_2^{f2f}=1\right),$$

which is indicated by shaded area B in Figure 5.1. These potential outcomes were imputed as part of the estimation procedure under the MAR assumption discussed above.

5.3.3 Estimation of single- and mixed-mode effects against the HMB

As discussed in section 5.2, we chose to use the web mode as a measurement benchmark for the HMB case, while keeping F2F as selection benchmark. Likewise the SMB, now web measurements substitute the true score, but the selection mechanism of F2F is still deemed optimal, so that the population parameter μ is substituted by the potential outcome $E(Y^{web}\middle|S^{f2f}=1)$. From the case study data, web measurements were available in the mixed-mode sample. Figure 5.2 demonstrates the location of these potential outcomes as shaded area C (the figure omits nonrespondents in the SMB sample and double-nonrespondents in the mixed-mode samples).

To estimate single-mode selection- and measurement effects of the telephone sample against the HMB, the potential outcomes $E(Y^{web}|S^{tel}=1)$ are required (shaded area D). Figure 5.2 demonstrates that these outcomes are counterpart to the potential outcomes $E\left(Y_1^{f2f}\middle|S^{tel}=1\right)$ required for the SMB case (area A). Finally, to estimate the mixed-mode effects the benchmark mean for follow-up respondents needs to be estimated (area E) analogous to area B in the SMB case. As in the SMB case, multiple imputation was applied to estimate these potential outcomes using the case study data. The details of this procedure are provided next.

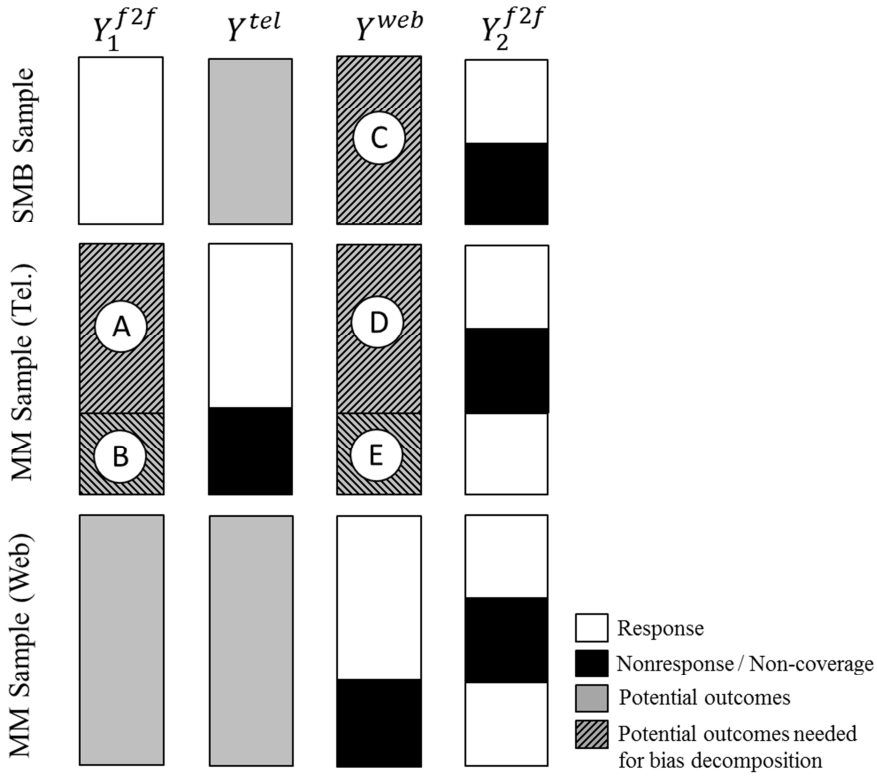


Figure 5.2: Missing data pattern of an extended within-subject design with two different mixed-mode samples (telephone and web) for use in the HMB case (nonrespondents in the SMB sample and double-nonrespondents in the mixed-mode samples omitted)

5.3.4 Practical implementation of the multiple imputation procedure

In the context of the missing data pattern shown in Figure 5.2, auxiliary information Y_2^{f2f} is used to impute potential outcomes, but Y_2^{f2f} itself suffered from unit nonresponse (cf. re-interview response rates, Table 5.1). Both missing data problems may be solved simultaneously by multiple imputation under the MAR assumption discussed in section 5.3.2 (Rubin, 1987; Schafer & Graham, 2002; Schafer, 1997). A sequential regression approach was applied, called multiple imputation by chained equations, using the software '*MICE*' (Raghunathan, Lepkowski, van Hoewyk, & Solenberger, 2001; Rubin, 2003; van Buuren, Brand, Groothuis-Oudshoorn, & Rubin, 2006; van Buuren & Groothuis-Oudshoorn, 2011; van Buuren, 2012, p. 109). In total, thirty different variables were imputed per mode concerning different aspects of the CVS, in particular the social quality and problems of the neighbourhood, frequency of

insecurity feelings, police contact and evaluation, and victimization[7] (see overview in appendix D).

In *MICE*, prediction models with an appropriate link function for each imputed variable were specified. Since the CVS variables were measured on polytomous, dichotomous, or interval scales, we applied multinomial, logistic and normal regression models, respectively[8]. A crucial element of the multivariate imputation is the selection of predictor variables. Given the large number of possible predictors in the data set, cautioning of over-specified models was important. To restrict the number of predictors, we applied the following procedure:

- Mode-specific potential outcomes of *Y* at occasion 1 were predicted by their repeated measure Y_2^{f2f}.
- Y_2^{f2f} were predicted by their four counterparts in the modes at the first occasion.
- Any Y_2^{f2f} variable from the follow-up exceeding a medium bivariate association (Cramér's *V*>.30) with a *Y* variable at the first occasion (e.g. Y^{tel}) was included. It should be emphasized that conditioning on these additional Y_2^{f2f} variables can further strengthen the MAR assumption.
- Eight socio-demographic background characteristics were available as additional predictors from the national population register for all units[9]. Any of these background characteristics exceeding a small association of *V*>.15 was included as predictor for any *Y* variable.

Fifty data sets were multiply imputed. The proportion of missing data was high in the present study, due to the fact that potential outcomes were imputed across four samples. However, the fraction of missing information, a model-based estimate of missingness (Rubin, 1987; van Buuren, 2012, p. 41), was below 50% in most cases. This fraction

[7] Victimization is surveyed on multiple variables and aggregated to two summary scores (past victimization in the last year [yes / no] and a count of victimization incidences).

[8] For the ordered polytomous scales (e.g., rating scales) proportional odds models could be used instead. These often led to convergence problems of the model fitting algorithm. In this case, MICE internally applies multinomial regression instead (van Buuren, 2012, p. 76). Since the problem occurred regularly in our data, we chose to model all variables directly by multinomial regression models.

[9] I.e., gender, age, income, household-size, civil status, nationality, urbanization of living area, and living in one of three large Dutch cities. The income variable shows a small amount of missing cases, which were treated as category in the present analyses.

suggests that when using fifty multiply imputed datasets, estimates of total variance are precise[10]. To estimate the within-imputation variances, all effects were bootstrapped by 1,000 iterated draws within each imputed data set. The within- and between imputation variances were pooled to the total variance using Rubin's rules (Schafer, 1997, pp. 109–110). Two sided significance tests were executed using t-tests with adjusted degrees of freedom.

Another relevant issue in estimation was the treatment of item nonresponse due to "don't know" (DK) or refused answering. In principle, there is not a unique way of handling this problem, because DK answers may be either regarded as missing information or substantial answer (i.e., as an answering category). The results presented are based on imputed item nonresponse as part of the missing data correction. Results turned out robust, because treating DK as answering category yielded very similar findings to the results presented here.

5.4 Results

In this section, we answer RQs 1 and 2 for the case of the CVS experiment, discussing the F2F SMB and the web/F2F HMB separately. We note again that these analyses are based on different benchmark assumptions and research interests, as exposed in the previous sections. After presenting results, we discuss implications for the CVS under both perspectives.

5.4.1 Effects against the F2F single-mode benchmark

RQ 1 requires estimating the total effect of using a different focal mode (web, mail, telephone) than the F2F benchmark and assessing the impact of the mixed-mode design on the single-mode effect. Twenty of the thirty included CVS variables were measured

[10] The fraction of missing information determines the strength of inflation of total variance estimates by the equation: $T_m = (1 + \frac{\gamma_0}{m})T_\infty$, where T_∞ is the ideal variance under infinite imputations, m is the number of imputations and $\gamma_0$is the fraction of missing information (Rubin, 1987, p. 114; van Buuren, 2012, p. 49). For a fraction of missing information of 50%, 50 imputations lead to approx. 0.5% higher standard error estimates than under the ideal situation (1% higher variance). The maximum fraction of missing information across all variables, modes, and effects was .771 for the case of the mixed-mode measurement effect of mail in the HMB case. Also in this case the total variance estimate is still precise (approx. 0.77% higher standard error estimates).

on polytomous answering scales and dichotomized similar to reporting by Statistics Netherlands (cf. appendix D for an overview). The dichotomized target statistic is a proportion (e.g., the proportion "agree or completely agree"). Four further variables are summary scales measured on interval level, and one represents a count of victimizations in the past year.

Table 5.2 presents a summary of the significant single- and mixed-mode effects on these variables. A majority of variables showed total effects against F2F in web (20) or mail (16), but fewer variables were affected in telephone (7). The counts beneath these numbers inform about the impact of the mixed-mode follow-up. We distinguish four possibilities: a mixed-mode effect is insignificant after the follow-up ($TB^{MM} = 0$) or it is still significant and, if it is significant, it may decrease significantly[11] ($|TB^{SM}| > |TB^{MM}|$), may stay equal ($|TB^{SM}| = |TB^{MM}|$) or may increase ($|TB^{SM}| < |TB^{MM}|$). It can be seen that for web and mail the follow-up was very effective, decreasing total bias against F2F in 18 (of 20) and 12 (of 16) cases, respectively. Half of these cases even reached insignificant level after the follow-up.

Table 5.2 does not allow an assessment of the size of effects, however. For this purpose, and answering RQ 2 below, we employ three scatterplots of single-mode against mixed-mode effects showing estimates for the 25 variables available on dichotomous scale (Figure 5.3, upper row). The lower row of scatterplots presents t-statistics for each variable and may be evaluated against critical values from the t-distribution. Critical values for a two-sided test ($p<.05$) are provided by horizontal (single-mode) and vertical (mixed-mode) dashed lines[12]. The diagonal line in all plots has slope one (i.e., it is not a regression line). Deviations from the line, therefore, imply change in effects by the follow-up.

---

[11] To evaluate significance of change between single and mixed-mode total effects, we bootstrapped the statistic $\Delta = TB^{SM} - TB^{MM}$ and tested its difference from zero. If the test was significant, we inspected, whether bias was increased or decreased relative to the benchmark, by evaluating $\left|T\hat{B}^{SM}\right| < \left|T\hat{B}^{MM}\right|$ against $\left|T\hat{B}^{SM}\right| > \left|T\hat{B}^{MM}\right|$ for the final effect estimates.

[12] T-tests of a zero-effect null hypotheses based on multiply imputed data require a variable-specific adjustment of degrees of freedom that is based on the between and within-imputation variance (Rubin, 1987; Schafer, 1997, p. 110). For this reason, all tests involve variable-specific critical values. For the present data the maximum critical value across all variables and modes is depicted as dashed line (approx. +/-1.975), which, however, did only deviate marginally from critical values of large sample t-tests (i.e. normal distribution critical values, +/-1.965, $p<.05$).

Table 5.2: Count of significant and non-significant single-mode total effects and the change induced by the F2F follow-up (mixed-mode effects) for the SMB and the HMB case (significance tests on p<.05)

Single-Mode	Mixed-Mode	F2F Benchmark (SMB)			Web/F2F Benchmark (HMB)			
		Web	Mail	Tel.	Web	Mail	Tel.	F2F
$TB^{SM} = 0$		10	14	23	30	20	9	7
	$TB^{MM} = 0$	9	13	22	20	18	9	-
	$TB^{MM} \neq 0$	1	1	1	10	2	0	-
$TB^{SM} \neq 0$		20	16	7	0	10	21	23
	$TB^{MM} = 0$	9	4	1	0	2	1	-
	$\lvert TB^{SM}\rvert > \lvert TB^{MM}\rvert \neq 0$	9	8	0	0	0	3	-
	$\lvert TB^{SM}\rvert = \lvert TB^{MM}\rvert \neq 0$	2	4	6	0	4	17	-
	$\lvert TB^{SM}\rvert < \lvert TB^{MM}\rvert \neq 0$	0	0	0	0	4	0	-
Total		30	30	30	30	30	30	30

Consider first the plot of single- against mixed-mode total effects (upper left hand). Single-mode total effects of web and mail are substantially larger than for telephone in many cases. Moreover, the seven significant single-mode total effects for telephone (Table 5.2) are found to be of smaller magnitude than for mail and web. Secondly, the impact of the mixed-mode follow-up is apparent for both web and mail, but not telephone, as estimates are moved towards zero mixed-mode effects (i.e., the horizontal axis). This effect is particularly pronounced for web suggesting that the mode profits more strongly from the F2F follow-up in reducing total effects.

RQ 2 asks about the sources of the total effects we identified. This question is addressed by the middle (selection effects) and right plots (measurement effects). It is immediately clear that selection effects were very small, and that measurement effects were the dominant component in creating effects between the SMB and the three focal modes. It is important to emphasize that the reduction in measurement effects by the F2F follow-up seemed to be effective, because the follow-up is conducted in the benchmark mode. The follow-up mode measurement effect (formula (5.10)) was indeed small in all cases (not shown here). The reduction in single-mode measurement effects is, therefore, dominated by the term $\hat{\pi} M\hat{B}^{SM}$ (formula (5.9)). The web mixed-mode sample showed a substantially higher amount (51.4%) of F2F follow-up respondents

suggesting that single-mode measurement effects were, roughly, reduced by this factor, whereas mail and telephone were impacted less strongly (26.7%, 25.4%; cf. Table 5.1).

5.4.2 Effects against the hybrid-mode benchmark

The HMB takes web measurements as benchmark while allowing for the selection mechanism of F2F. Significance tests of the total effects against the HMB are provided on the right side of Table 5.2. In addition to the three mixed-mode designs, effects for the single-mode F2F survey are shown against the HMB. It is apparent that for telephone and F2F, a large number of variables (21 and 23, respectively) showed significant total mode effects, whereas for mail fewer variables (10) reached significant level. There were no significant total effects for web. With respect to web, it should be noted that only single-mode selection effects could have caused a total effect, given that web measurements are used as benchmark.

The F2F follow-up to the three modes was mainly ineffective or harmful. For web, it was even very harmful, increasing total effects in 10 cases. Similarly in mail, it increased total effects on 6 variables, while only reducing it on 2. It was mainly ineffective to reduce the telephone total effect against the hybrid web-F2F benchmark.

For a more detailed picture, we consider the scatterplots of the three single- and mixed-mode effects for the HMB case (Figure 5.4). Since the selection benchmark did not change, selection effects against F2F were small and insignificant, likewise the SMB case. However, telephone showed large measurement effects against the web measurement benchmark and the F2F follow-up was ineffective. Mail showed smaller single-mode measurement effects against web or no effects. However, it can be seen that the F2F follow-up increased measurement and total effects on many mail variables (cf. points below the diagonal line) reflecting that the F2F follow-up was not beneficial to the mail single-mode bias.

In addition, it can be seen that web, as measurement benchmark, does not exhibit single-mode measurement effects. Measurement effects are caused by the F2F follow-up, however, reflecting that F2F measurements showed a bias against the web benchmark. For this reason, the mixed-mode total effect of web is determined by the

follow-up measurement effect of F2F against the HMB (i.e., $(1-\hat{\pi})M\hat{B}^{FU} = .514M\hat{B}^{FU}$).

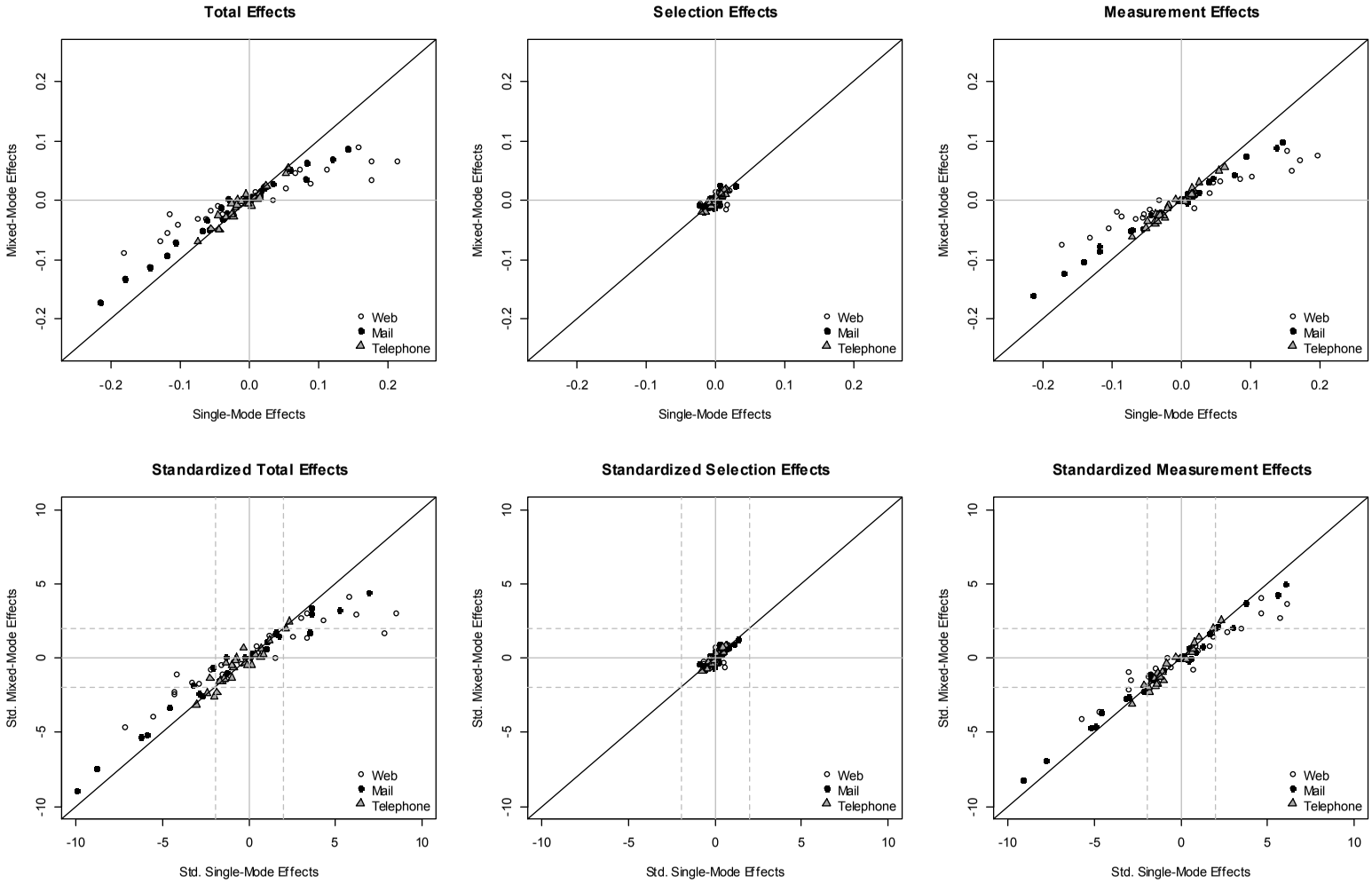


Figure 5.3: Scatterplots of single-mode against mixed-mode effects for the SMB case (upper row: unstandardized effects; lower row: standardized effects, where dashed lines indicate critical values [p<.05])

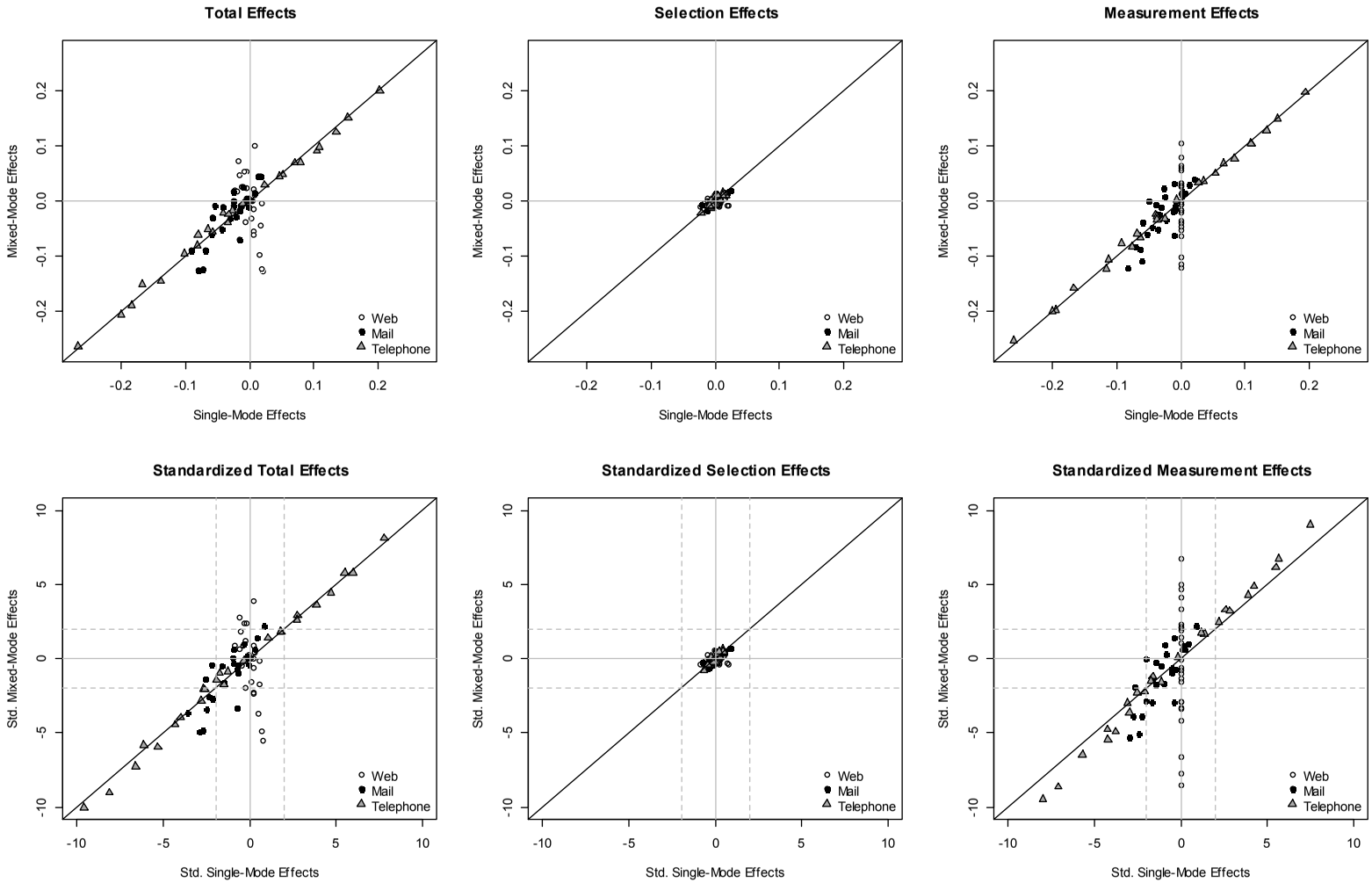


Figure 5.4: Scatterplots of single-mode against mixed-mode effects for the HMB case (upper row: unstandardized effects; lower row: standardized effects, where dashed lines indicate critical values [p<.05])

5.4.3 Evaluation of effects by variable groups

Finally, we consider measurement and total effects by the type of variables and statistics reported in the above analyses (Table 5.3). Significance of measurement and total effects is evaluated separately. It can be seen that for some cases a total effect did not imply a measurement effect (e.g., 11 of 17 significant web total effects imply a measurement effect). For these cases, a clear conclusion about the source of total bias against the benchmark cannot be drawn (i.e., a selection effect may represent an alternative explanation). However, for the majority of variables, a total effect did imply a measurement effect, reflecting the observations from Figure 5.3 and Figure 5.4. Also considering the clear measurement effect pattern of both figures, it is plausible that measurement effects underlay also the insignificant cases, where a total effect is observed but remains within observable differences.

We found that, regardless of mode and benchmark, all CVS variables may have been subject to a measurement effect, regardless of type of variable and answering scale. In particular the two groups of questions on the social quality and problems of the neighbourhood appeared susceptible to measurement effects. However, not all questions of any given group show clear measurement (or total) effects. We may conclude that measurement effects appear to be a general, but still a question-dependent phenomenon in the CVS. However, the strong presence of web and Mail measurement effects against F2F, and telephone and F2F effects against the HMB (web measurements) is again evident.

The only characteristics which were not affected by measurement across modes were two ‘victimization’ variables (victim of crime past 12 months / count of victimization), which represent key statistics in the CVS. The insensitivity to effects is generally a positive result. However, it should be noted that not all standard victimization questions could be included in the re-interview design and the two index variables are based on shortened versions of the standard statistic. For this reason, the present findings about victimization should be interpreted with care.

Table 5.3: Count of significant single-mode measurement / total effect estimates against the SMB and HMB by variable types (significance tests on p<.05)

		F2F Benchmark (SMB) (Sig. MB^{SM} / TB^{SM})			Web/F2F Benchmark (HMB) (Sig. MB^{SM} / TB^{SM})			
	No. of items	Web	Mail	Tel.	Web	Mail	Tel.	F2F
Proportion statistics								
Social quality[a]	9	5 / 6	8/ 8	2 / 2	0 / 0	2 / 2	7 / 6	7 / 7
Neighbourhood problems[b]	8	6 / 7	4/ 4	1 / 2	0 / 0	5 / 6	8 / 8	7 / 7
Insecurity feelings[c]	4	0 / 2	1/ 1	0 / 0	0 / 0	0 / 0	3 / 2	2 / 2
Police contact (yes / no)	1	0 / 1	0 / 0	0 / 1	0 / 0	0 / 0	0 / 0	0 / 0
Police evaluations[d]	2	0 / 1	0 / 1	0 / 1	0 / 0	0 / 0	1 / 1	1 / 1
Victim of crime[e]	1	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0
Total	25	11 / 17	13 / 14	3 / 6	0 / 0	7 / 8	19 / 17	17 / 17
Mean statistics								
No. of victimizations[f]	1	0/ 0	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0
Quality of life rating[g]	1	0/ 0	1 / 0	0 / 0	0 / 0	0 / 0	0 / 1	0 / 0
Neighbourhood scales[h]	3	2 / 3	2 / 2	1 / 1	0 / 0	1 / 2	3 / 3	3 / 3
Total (incl. dichotomous)	30	13 / 20	16 / 16	4 / 7	0 / 0	8 / 10	22 / 21	20 / 20

For full details on all items see Appendix D.
a. Likert scale items: % (completely) agree (5 answering categories)
b. Frequency scale items: % (frequently or sometimes) (3 answering categories)
c. Insecurity feelings: 2 items (% yes), 2 items (% frequently or sometimes)
d. % very satisfied or satisfied
e. % victim in the past 12 months (aggregated across multiple items)
f. Count of victimization past 12 months (aggregated across multiple items)
g. Score on 10-point scale from very low (1) to very high (10)
h. Aggregated summary indices based on multiple social quality and neighbourhood problems items.

5.4.4 Conclusions for the CVS

In drawing conclusions for the CVS, it is important to recall the objectives of the SMB and HMB. In the SMB case, both measurement and selection mechanisms are taken from the same mode (F2F, in this study). F2F can be considered a historical benchmark for the CVS, which was a F2F survey upon first introduction. This study revealed that when using the web or the mail mode instead of F2F, it is impossible to avoid a strong change in statistics for a large number of CVS variables (RQ 1). The reason for these effects was an increase in measurement bias in web and mail relative to F2F. For telephone, measurement effects were also present, but on much smaller scale and in smaller number. In conclusion, when F2F is the desired benchmark estimate, use of

single-mode web and mail should be avoided. Use of telephone may be viable when accepting some smaller systematic changes.

The classical motivation of a sequential F2F follow-up is reducing single-mode selection bias (RQ 2). Selection effects against F2F, however, were not identified on a statistically significant level in this study. Still, estimates from the web and mail mixed-mode surveys were often closer to the SMB than the single-mode estimate alone, because the F2F follow-up provided measurements that were very similar to the single-mode F2F benchmark, as can be expected. To achieve this effect, a F2F follow-up would be chiefly desirable for both modes, but not for reducing single-mode selection effects. However, there remain a number of relatively large total effects suggesting that this procedure cannot fully compensate the measurement bias created by web and mail. The response proportion π determines the strength of follow-up mode impact. In future research it is important to evaluate the mix of biases and the role of π further.

The objective of the HMB is to optimize a benchmark with respect to both measurements and selection mechanisms. Under the assumption that web measurements are superior to F2F, e.g. due to anonymous answering, we used web as measurement benchmark instead of F2F. This change strongly affects conclusions about the CVS. Since we did not identify any selection effects against F2F, the optimal mode would now be a single-mode web survey. Furthermore, using telephone or F2F would suggest increasing measurement bias and should be avoided. However, the mail mode showed only smaller effects against the HMB. These were primarily limited to a single group of questions ('neighbourhood problems', Table 5.3). In many cases, mail may, therefore, evoke similar estimates as web. However, for both mail and web, the F2F follow-up would only introduce measurement bias. Therefore, a mixed-mode design involving F2F should be avoided. Moreover, given these findings web and mail may be compatible for sequential mixed-mode surveys themselves. In the absence of selection effects, this may seem unnecessary. However, web yielded a small response rate (29.0%), which may be raised by including a mail follow-up, for example (Millar & Dillman, 2011). The expected mode effects in a sequential design of web and mail could not be evaluated in the present study, so that this advice is speculative.

In sum, these results reflect theoretical and empirical arguments in the literature that self-administered and interviewer modes often form a dichotomy with respect to measurement bias (De Leeuw, 1992, 2008; Klausch et al., 2013b), but also suggest that there may be exceptions to the rule and mode effects remain question dependent phenomena. A question-specific evaluation of effects is, therefore, necessary for any mixed-mode survey. In the CVS, it was eventually decided for a non-interviewer mode only redesign based on these results and practical considerations.

5.5 Discussion

Evaluating total bias, measurement bias, and selection bias, of mixed-mode designs before their introduction or redesign is a problem of great concern (RQ 1 and 2). In the absence of true scores, we suggested using measurements and selection mechanisms as benchmark, which are defined as optimal yielding either single-mode (SMB) or hybrid-mode benchmarks (HMB). It is important to distinguish between single-mode and mixed-mode effects against the SMB and HMB. Evaluating single-mode effects is relevant to assess the need for mixed-mode designs to reduce bias, while mixed-mode effects inform about the success of the procedure. In doing so, selection and measurement effects indicate the sources of the observed total mode effects.

Total mode effects may be estimated from simple comparative designs in the SMB case. However, evaluating measurement and selection effects, as well as effects against a HMB, requires estimating potential outcomes. This objective necessitates additional auxiliary data. We suggested using re-interview data for this purpose. Our design is related to other re-interview methods for bias estimation including the basic question approach (Kersten & Bethlehem, 1984), the call-back approach (Elliott et al., 2000; Hansen & Hurwitz, 1946; Keeter, Miller, Kohut, Groves, & Presser, 2000), and test-retest designs (Biemer & Lyberg, 2003, p. 291), but also features important differences. The basic question and call-back approach involve re-interviews of nonrespondents to estimate nonresponse bias. Test-retest designs normally aim at estimating measurement error. In doing so, measurement independence is often assumed. Our design allows estimating both measurement and selection effects against benchmarks. Furthermore, by modelling potential outcomes at the first occasion, we explicitly allowed for the

possibility that re-interview measurements and selection mechanisms can change between initial interview and follow-up. Change may occur, if follow-up F2F respondents provide different answers than in the benchmark mode (e.g., due to experienced response burden). An alternative explanation is substantial change in statistics across time, but it is often likely to be small in the study period of sequential mixed-mode surveys (two months, in the present study).

An advantage of the design is that it is tailored for use in parallel to sequential mixed-mode surveys and can be implemented even for ongoing surveys without affecting the standard fieldwork (i.e., the benchmark sample is independent and the re-interview does not impact the standard mixed-mode fieldwork). Furthermore, our design does not require the follow-up to be conducted in the measurement and selection benchmark mode. Although in the current case study F2F served as, both, the SMB and follow-up mode, estimation would still be possible with a different follow-up or benchmark mode. For example, a design using telephone as follow-up to web, while F2F remains the SMB, can also be evaluated, if the re-interview is conducted in telephone.

Another advantage of our design is that it allows for a structured view on design decisions in mixed-mode surveys. We demonstrated how single- and mixed-mode designs can be evaluated against SMB and HMB and how a top-down evaluation can be performed (i.e., from total mode effects to its selection and measurement components). Such a top-down approach supports data collection and questionnaire designers. A decisive role is played by the size of the experimental samples and effect sizes that are required to be observable. Future research should evaluate minimum mode-specific sample size requirements, also considering costs of the experimental design.

Our results and design should be judged against a number of limitations which show up paths for further research. Firstly, the size of effects should be evaluated against costs and budgets. For example, F2F is the most expensive mode in data collection, whereas web is inexpensive. The web-F2F mixed-mode design could reduce bias compared to single-mode web (SMB case), but the F2F follow-up response (cf. Table 5.1) may cause substantial additional costs over single-mode web. Further

research, therefore, needs to weigh off mode effects against budget constraints (Vannieuwenhuyze, 2014).

Secondly, the re-interview measurements, Y_2^{f2f}, are assumed to be observations from the same response distribution regardless of sample assigned at the first occasion (e.g., F2F benchmark or mixed-mode telephone) or being respondent or nonrespondent in these samples. The first part of this assumption was checked by comparing response distributions across assignment conditions (Klausch et al., 2014). We did not find any strong evidence for dependence in the present study[13]. The second part suggests that follow-up respondents in the mixed-mode design provide equivalent Y_2^{f2f} measurements as re-interviewed respondents. In the present study, a relatively long period (6-8 weeks) lay in between first interview and re-interview, which may turn this assumption plausible. However, developing approaches for evaluating this assumption is necessary in the future, especially for use in designs which allow less time between first and second stage of the sequential mixed-mode design.

Thirdly, our results may depend on the specific estimation procedure (multiple imputation) and related modelling decisions. For example, we included Y_2^{f2f} as predictors of potential outcomes and it might be an option to extend the models to observed information at the first interview (e.g., Y^{tel}). We started to evaluate this possibility but including such predictors led to convergence problems of the Gibbs sampler in *MICE*. We suspect this may be related to either the fact that the partial correlation between response distributions at the first occasion is not observed or to the discrete level of measurement of many variables leading to highly parameterized models. Future research could, therefore, evaluate more general imputation and modelling strategies for the missing data pattern of the design for both continuous and

[13] We conducted chi-square tests of independence between the mode assignment indicator and the polytomous items of the re-interview. For the scaled variables, log-likelihood ratio tests of an empty linear regression model against a model including the mode assignment indicator as predictor were employed. Rao-Scott adjusted tests accounting for sampling weights were applied (Lumley, 2010). We found that 3 out of 30 variables showed significant differences across sample assignment under $p<.05$. This number of significant tests may occur by chance due to multiple testing. After Bonferroni adjustment for multiple testing these differences disappeared. Furthermore, empirical differences in distributions were very small.

discrete data and the use of alternative imputation techniques (e.g., predictive mean matching; van Buuren, 2012, pp. 68–74).

The evaluation of bias in mixed-mode surveys will continue to be of concern for methodologists and practitioners. In this respect, further development of our method is desirable. If the strong prevalence of measurement effects should prove a problem in many mixed-mode surveys, developing adjustment methodology for measurement effects is necessary. Such methodology could use Bayesian approaches that yield multiple imputations of potential outcomes in mixed-mode surveys. Developing and evaluating these methods appears urgent in face of our empirical results.

6 Discussion and Outlook

This dissertation dealt with methodology for evaluating three central objectives during the design of mixed-mode surveys, which combine modes of data collection at the administration stage (cf. chapter 1). These objectives consider the extent of selection error accumulated by single- and mixed-mode designs, measurement equivalence between modes of data administration, and the optimal trade-off between measurement and selection error in the total survey error of mixed-mode data collection designs. Evaluating these objectives is not trivial in practical situations when true scores are not available. An additional problem is caused by the confounding of measurement and selection effects in the total effect between survey modes in single- or mixed-mode designs.

Each of chapters 2 to 5 made a methodological contribution to evaluating the general objectives in practice, where methodological developments were influenced by the data basis of the MEPS experiment (cf. section 1.5.1). This data basis allowed a comparison of single modes of data administration (face-to-face, telephone, mail, and web) as well as evaluating the impact of a sequential mixed-mode follow-up (face-to-face). Empirical results on evaluations of the three objectives for the case of the Crime Victimization Survey (CVS) were presented and conclusions about mixed-mode redesign options for the CVS were drawn. In the next section, the empirical results of the previous chapters are summarized. The subsequent discussion addresses the question to which degree the presented methodological approaches can be generalized to evaluate survey errors for other surveys and types of mixed-mode designs.

6.1 Summary of Empirical Results

Objective 1 required an assessment of the size of selection error of single-mode designs and a comparison of the effect of a mixed-mode extension on the single-mode error (chapter 2). We found small differences in selection error on socio-demographic characteristics between single-mode surveys when compared on an absolute level against the sampling frame. The face-to-face follow-up to web, mail or telephone clearly turned the socio-demographic profile of the mixed-mode response sample more similar to a single-mode face-to-face survey. For the CVS variables, absolute and relative selection error was very small and mainly insignificant, but there was some evidence that the follow-up could also turn absolute selection error on CVS variables more similar to face-to-face. These findings implied that selection error on some socio-demographic and CVS variables was increased and on others reduced, mostly to the level or in the direction of the selection error of a single-mode face-to-face survey. A general reduction of selection error against the sampling frame was not possible using the sequential designs under consideration (objective 1).

It therefore appeared that face-to-face follow-ups may be capable of turning the selection error profile, at least for socio-demographics, similar to a single-mode face-to-face survey. If this error is considered desirable, the mixed-mode survey would be useful (but a general reduction in error is not feasible). In addition, given the generally small effect sizes, it is questionable, whether the extra effort by a follow-up is justifiable in practice.

The predominant empirical phenomenon was represented by measurement effects on systematic and random measurement error, leading to measurement non-equivalence between survey modes (objective 2, chapters 3 to 5). The primary measurement effects were present between the interviewer and the self-administered modes, and as far as attitudinal rating scale questions were concerned, chapter 3 identified method-level sources of bias, which affected all items in a scale of questions equivalently. Besides bias, chapter 3 additionally studied error variance components, finding that modes differed with respect to the extent of random error on some CVS questions, where the two self-administered modes very often evoked smaller random error than the interviewer modes. There is therefore reason to believe that the presence of

measurement effects is primarily a property of modes and did not strongly depend on question content (cf. also the key-variable analysis in chapter 4). Chapter 5 extended the analysis to a larger set of CVS questions. Considerable variance in the strength of single-mode measurement effects was found across question, but the majority of effects were still located between the self- and the interviewer-administered modes.

Methodological research on designing mixed-mode questionnaires that prevent the occurrence of measurement effects, such as unified mode questionnaires or general mode questionnaire designs (De Leeuw, 2005), should therefore evaluate in more detail, whether it is at all possible to create measurement equivalence between interviewer and self-administered modes or if the differences found in this study are a symptom of non-avoidable differences in the measurement process (cf. theoretical discussions in chapter 1 and 3).

The trade-off between measurement and selection error (objective 3) was evaluated against a single-mode face-to-face benchmark or a hybrid-mode web/face-to-face benchmark (chapter 5). Selection effects against the face-to-face selection benchmark were absent in this analysis and also a change in selection error by the follow-up (mixed-mode effect) was not identified. Therefore, the trade-off was dominated by the measurement effect component, which was the primary cause of total mode effects (i.e., differences in TSE of single- and mixed-mode estimates against the face-to-face and web/face-to-face benchmarks). If face-to-face represented the measurement benchmark, for example, sequential designs or single-to-single mode switches involving telephone seemed still appropriate, but the self-administered modes often increased TSE of a mixed-mode design due to measurement effects. Conversely, when web, as a self-administered mode, was the measurement benchmark, these conclusions were all reversed (i.e., single-mode web and mail were the ideal mode choices, and follow-up by the face-to-face mode increased total/measurement effects). Methodological considerations about choosing the 'right' benchmark seem important for this reason, as discussed in more detail in the next section.

Given the centrality of measurement effects in the trade-off of selection and measurement, the importance of developing methodology for preventing measurement effects is further underlined. Another important line of research is represented by

methods to address measurement effects after they have occurred, for example by statistical adjustment (i.e., if measurement effects cannot be prevented). Further discussion on this important aspect is provided in the outlook (section 6.5).

In conclusion, it is suggested to be cautious about generalizing all findings to other survey topics, variables, and populations. The number of potential causes of mode effects on measurement and selection error (cf. literature reviews in chapter 1 to 3) gives reason to believe that for some survey topics and variables (e.g., surveys on sensitive topics, such as health conditions) selection effects might play a more dominant role and measurement effects may be different. From this point of view, the empirical results may be specific to the CVS and the MEPS experiment and replications under different circumstances are very desirable. To support researchers in this endeavour the following three sections discuss the most important steps, fallacies, and considerations in implementing similar research designs for evaluating mode effects on total survey error, measurement error, and selection error in other surveys (sections 6.2 to 6.4).

6.2 The Importance of Selecting Benchmark Modes for Evaluating Systematic Survey Error

The absence of true scores in the practice of survey research is an inherent problem in studying total survey error (TSE) and its components. The choice of a benchmark mode for both measurements obtained in a survey as well as the selection mechanism is, therefore, an important step to interpret results from mode comparisons. A benchmark mode can have two components. If a mode is the ‘measurement benchmark’ its measurements substitute for unobserved true scores in the benchmark estimate (cf. chapter 5). The availability of benchmark measurements is crucial for the evaluation of systematic error components. Without the availability of benchmark measurements, a mode difference in estimates is relatively meaningless, denoting merely a difference in systematic TSE (cf. section 1.4.2). However, when one of the modes is set as a measurement benchmark, any measurement effect in a mixed-mode design can be interpreted as an increase in systematic measurement error. Furthermore, any selection effect denotes that ‘different types’ of respondents are reached by two different modes. If a mode is additionally set as a ‘selection benchmark’ (e.g., single-mode face-to-face)

its selection mechanism in a survey design is viewed as 'optimal' or 'desirable' for the survey target variables at hand. A selection effect then denotes an increase in selection error relative to the selection benchmark.

In this section, it is first reviewed which role the benchmark mode approach played across chapters 2 to 5. Subsequently, the approach is discussed in more general context to allow its application in different situations than the experimental design applied in this study (MEPS experiment) and the related methods presented in the previous chapters.

6.2.1 Development of the benchmark mode approach in this dissertation

The importance of the benchmark mode became clear as methods were developed across chapters 2 to 5. In chapter 2, the face-to-face re-interview survey of the MEPS experiment was applied as measurement benchmark. Unit nonresponse in this set of 'benchmark data' was adjusted using auxiliary data from the register. The re-interview in face-to-face, therefore, did not serve as a selection benchmark, yet, but the measurements were used as benchmark to study selection error against the sampling frame. This situation may be most comparable to the record check approach, in which for all sample members (regardless of response and nonresponse) data from an external register are available (Biemer & Lyberg, 2003; Kreuter et al., 2008). As such the measurement benchmark in the re-interview approach presented in chapter 2 is similar to creating a list of records, which has to be extrapolated to the sampling frame due to missing data from unit nonresponse. It should be noted that also administrative records have, at one point, been created using a method of measurement (e.g., administrative questionnaires filled in upon registration at a mailing address with local authorities). Given that such administrative information may be erroneous (Bakker, 2012), using a re-interview approach for 'creating records' may not be much less plausible than using administrative records. In survey methodology, the use of re-interview designs for assessing survey error is not new (e.g., 'call-back approach'; cf. discussion section in chapter 5). However, the application of re-interview data in a mixed-mode context for solving the confounding problem of measurement and selection error represented a novel contribution.

In chapter 4, different conditional measurement effects were defined depending on choice of measurement benchmark mode and target mixed-mode design (Table 4.1 in chapter 4). The idea was developed that if one measurement is optimal, certain conditional measurement effects need to be negligible (zero) to guarantee measurement equivalence (in expectation) in sequential mixed-mode designs or mode-switch designs. Since the focus was set on assessing measurement equivalence, the selection benchmark was not named explicitly, however. In chapter 5, finally, measurement and selection benchmarks were introduced more formally. These ideas also allowed extending the available types of benchmarks from a single-mode survey (e.g., face-to-face) to a 'hybrid-mode' benchmark, where selection mechanisms and measurements were mixed. Also the difference between single-mode and mixed-mode effects against the benchmark was introduced, where single-mode effects can be compared to mixed-mode effects to assess the efficacy of a sequential mixed-mode design. These ideas are discussed in more general context in the next section.

A special role is played by the measurement equivalence analyses presented in chapter 3, in which marginal equivalence was tested without making any assumptions about measurement benchmarks. Such a procedure is possible, but requires strong research hypotheses to guide equivalence tests (e.g., to caution about over-fitting and capitalization on chance), as demonstrated in that chapter.

6.2.2 Formalization of the benchmark mode approach

Given the importance of the benchmark approach to identify systematic survey error, it is useful to consider its role in the context of the most important mixed-mode designs that combine modes of data collection in the administration phase and thus generalize the contributions of the previous chapters (Figure 6.1). In doing so, it is distinguished between three types of mixed-mode designs: sequential, concurrent (introduced as type III designs in chapter 1), and cross-sectional/longitudinal with a mode-switch between waves (type IV/V). For purposes of illustration, the sequential and concurrent designs are assumed to have three modes, but including more or less modes in the designs is possible. This schematic relates to many real world sequential designs that, for example, commence by a web survey, follow up nonrespondents by telephone and, finally, in

person (cf. examples on the American Community Survey and Crime Victimization Survey provided in section 1.2.3).

Additionally, a distinction is made between two situations: first, the benchmark mode is available as a single-mode survey (single-mode benchmark design [1]), and second, the benchmark mode is part of a mixed-mode design (sequential [2], concurrent [3] or mode switch designs [4]). In the first situation, an independent survey is assumed to be the selection and measurement benchmark, which is implemented for evaluation purposes (mode C is the 'Comparison / Benchmark Design' in Figure 6.1). In the second situation, one mode in the mixed-mode design is assumed to be the measurement benchmark and a single-mode benchmark survey is not necessarily available (in Figure 6.1, mode C is assumed to be the benchmark in mixed-mode designs 2 to 4). Against these benchmarks, different types of mode effects can be defined: single-mode effects, mixed-mode effects, follow-up effects, and relative mode effects. Each of these effects denotes a difference in (systematic) total survey error of an estimate from a response group to the benchmark (total mode effect). It consists out of a measurement and a selection effect against the benchmark (cf. chapter 5). Both situations are discussed in the following.

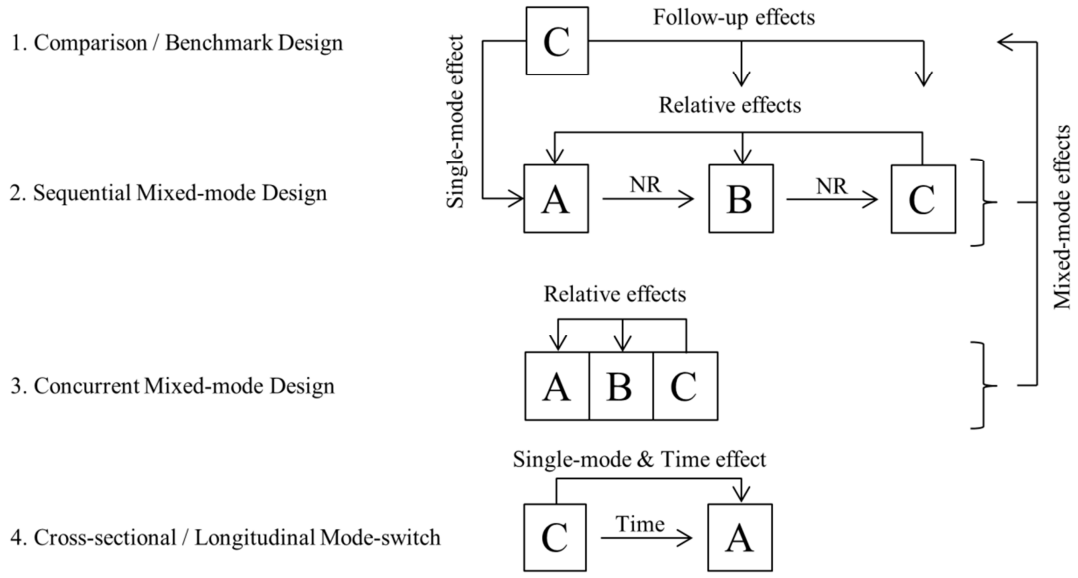


Figure 6.1: Schematic illustration of different types of mode effects evaluated against a benchmark mode C, which is defined either as part of the mixed-mode design (relative perspective) or observed as a single-mode comparison design (a box denotes a response group to a given mode; NR denotes nonresponse or non-coverage)

First, consider the situation when a single-mode benchmark survey C is available and compared to a sequential design (design 1 in Figure 6.1). As explained in detail in chapter 5, a single-mode effect denotes the difference in mean estimates between the initial mode of a sequential design and the benchmark (design 2). A mixed-mode effect denotes the total difference in mixture estimate from a mixed-mode survey (e.g., simple response mean across all response groups) and the single-mode benchmark. Finally, a follow-up effect denotes the difference between follow-up response group estimates (B or C) and the single-mode benchmark. The mixed-mode effect is a composite of single- and follow-up effects weighted by response group sizes in the mixed-mode design. The mixed-mode effect can be defined equivalently for a concurrent mixed-mode design (design 3). Now follow-up effects are absent, but each mode-specific response sample in the concurrent design also shows a mode effect against the benchmark (not shown in Figure 6.1).

Second, when the single-mode benchmark survey (C) is not available, instead a specific mode in the sequential or concurrent designs may be considered the benchmark for measurements. When effects are evaluated against this benchmark mode, relative mode effects are considered (Vannieuwenhuyze & Loosveldt, 2013). As shown in Figure 6.1, relative mode effects are defined between mode-specific response groups in sequential and concurrent mixed-mode designs (2 and 3). Contrary to the single-mode benchmark, it makes less sense to set the measurement benchmark mode in a mixed-mode design also as the selection benchmark, because it may represent a selective group of respondents (e.g., in the sequential design, mode C denotes the response groups after selection in mode A and B). This disadvantage of the relative perspective is discussed in more detail in the next section

In the case of a mode-switch design (design 4) from the benchmark (C) to a different mode (A), the mixed-mode effect is equivalent to a single-mode effect (Figure 6.1, last row), but the effect observed in the mode-switch design may be confounded with a time effect (i.e., substantial change across time). Evaluating mode switches may, therefore, rely on comparisons of modes administered at the same point in time to different samples or when assuming time-stability in time series data (cf. chapter 4).

In the discussion above, a hybrid-mode benchmark (introduced in chapter 5) is not considered, but it can be included in a research design for estimating mode effects. Some details for doing so are discussed in chapter 5, but generalized research designs are subject to further research.

6.2.3 Using mode effect estimates in practice

An important question is how to use mode effects in the practice of survey design, especially with regard to research objectives 1 to 3. First of all, it needs to be understood that the objectives require assessing measurement and selection effects besides total mode effects. In order to be able to judge the size of measurement and selection effects, a decomposition of the total mode effects needs to be possible. For the moment, it is assumed such decompositions are possible (further discussion on decomposing the effects is provided in section 6.3).

Total relative mode effects indicate the change in survey error against the benchmark mode within a mixed-mode design. In general, such an effect may be positive. If the total relative mode effect is caused by a relative selection effect, it is indicated that different modes might reach different 'types' of respondents (relative selection effects). This is one goal of concurrent or sequential mixed-mode data collection. However, relative measurement effects represent an alternative explanation and suggest that an increase in measurement error is caused by the mode switch – an undesirable effect. Therefore, the balance of relative selection and measurement effects in the total relative mode effect provides an indication on the success of the mixed-mode design (objective 3). Moreover, relative measurement effects of particular response groups in the design (i.e., conditional measurement effects) may be indicative for the need to improve questionnaire designs in order to equalize benchmark measurements (i.e., in Figure 6.1, measurements from mode C) and measurements under other modes (objective 2).

A disadvantage of the relative perspective is that relative selection effects have no normative meaning (chapter 5). Relative selection effects do indicate that mixed-mode designs reach different types of respondents with respect to a target variable, but it is unknown, whether selection error is increased or decreased by the design (objectives 1

and 3). A second disadvantage of the relative perspective is that follow-up mode measurements in a sequential design (follow-up mode C in Figure 6.1) may not represent appropriate measurement benchmarks, because the group of respondents from which benchmark measurements are taken is a specific and selective sub-population of late respondents, which may provide larger measurement error than usual respondents in the benchmark mode.

These shortcomings are addressed by setting the selection and measurement benchmark using another survey design such as the single-mode benchmark (comparison / benchmark survey in Figure 6.1). Under this perspective, the single-mode effects denote the 'starting error' of a sequential mixed-mode design (e.g., total single-mode effect of mode A against benchmark C in Figure 6.1). It is usually the objective to reduce this error by the mixed-mode follow-ups, since in all other cases (errors stay equal or are increased) the sequential design would not be needed. The single-mode effect is compared against the total mixed-mode effect to judge, whether such a procedure 'overall' was successful (objective 3). Another advantage of this perspective is that selection effects receive a clearer definition and interpretation. Moreover, single-mode and follow-up measurement effects are informative about the degree of measurement error accumulated by the design. Again such effects might be prevented by appropriate questionnaire design, where the focus is now either set on nullifying single-mode or follow-up measurement effects to achieve measurement equivalence to the benchmark (objective 2).

In concurrent designs, such an elaborate stepwise procedure is not possible (all modes are offered to respondents directly). However, evaluating different estimates from single-mode response proportions of the design or any mixture estimates thereof may provide hints on the optimal mixed-mode effect. Developing a formal approach for this procedure is still open to research.

In single-mode switch situations, the analysis is somewhat simpler focusing the estimation of single-mode effects (cf. chapter 4). It should be noted that assuring measurement equivalence by appropriate mixed-mode questionnaire design may not be sufficient to avoid breaks in time series, if single-mode selection effects are present. To

avoid selection effects adaptations to field work procedures may be necessary besides assuring measurement equivalence.

6.2.4 Choice of benchmarks

A crucial question is which factors should influence the choice of measurement and selection benchmarks in practice. This choice is non-trivial and represents an issue for further research (cf. section 6.5). So far, several general guidelines for this decision may be suggested.

The first decision to be made is that of using a mode within a mixed-mode design as a benchmark (relative perspective, cf. Figure 6.1) or using an independent survey design as a benchmark (single-mode benchmark, hybrid-mode benchmark). The first step in this choice is considering the status quo of the survey. If an ongoing single-mode data collection design is switched to a sequential or concurrent mixed-mode design or different single-mode design, the consistency of time series data (i.e., absence of structural breaks) is an important concern. Confounding of substantial change across time with changes in the TSE of estimates from the mixed-mode design is to be avoided. Setting the single-mode design, which has provided data for the time-series, as the single-mode benchmark thus is a natural choice. For example, in a repeated cross-sectional survey using face-to-face, a change to a sequential mixed-mode design of web followed by telephone should avoid all changes in TSE against the face-to-face time series. Using face-to-face as single-mode benchmark, consequently, is useful.

A second consideration about the choice of benchmark is costs. The implementation costs of an experimental comparison of designs, which is required for a single-mode (or hybrid-mode) benchmark can be quite substantial (and collecting re-interview data for disentangling measurement and selection effects can cause additional costs, as discussed in section 6.3). Therefore, a pragmatic choice may be focussing on relative mode effects instead of single-, follow-up, and mixed-mode effects (cf. Figure 6.1). The advantages and disadvantages of this perspective were discussed in the previous section. In particular, in the relative perspective, selection effects receive only a relative interpretation, and the measurement benchmark may need to be taken from a selective group of late respondents in sequential designs. Such respondents may provide larger

measurement error bias in the benchmark mode than expected from respondents in a single-mode survey administered in the measurement benchmark mode.

These considerations should further involve, whether there is a known mode of administration which provides the 'preferred measurements' and/or 'preferred selection mechanism' for the survey variables at hand. A measurement benchmark should be generally a mode which is trusted in terms of the quality of the administration process. General questionnaire design guidelines and plausibility arguments may support researchers in this choice (De Leeuw, Hox, & Dillman, 2008; Dillman, Smyth, et al., 2009; Groves et al., 2010). For example, interviewer modes may represent bad measurement benchmarks for socially desirable questions due to the threat for social desirability bias. If a relative perspective is chosen it should be plausible, furthermore, that measurements from follow-up respondents (e.g., face-to-face in a sequential design) can still be regarded of measurement benchmark quality. Clearly, the relative perspective can never be chosen, furthermore, when the preferred measurement mode is not part of the mixed-mode design.

Finally, setting additionally a selection benchmark implies that the selection mechanism evoked by a mode-specific fieldwork design is assumed optimal. As such it need not necessarily be provided by the realization of a single-mode survey, but could also be provided by any other (mixed-mode) design that is assumed to evoke the best selection mechanism with respect to the target variables. In chapter 5 it was argued that response and coverage rates are important indicators for a selection benchmark. It can be added that, under the 'best' selection mechanism, units are missing completely at random (MCAR) with respect to a target variable (Rubin, 1976). This situation is given when response propensities with respect to a target variable (or all target variables) are constant in the population (Schouten et al., 2009). Under this condition, the systematic selection error of a survey design would be negligible. Indicators for the representativeness of survey response (R-indicators) assess the variance of response propensities with respect to auxiliary data (id.; cf. chapter 2). Therefore, a selection benchmark could be chosen based on the survey design evoking the highest R-indicator in addition to considering response rates. Nevertheless, it is likely that no mode is ideal with respect to selection on all variables. For this reason, the choice of selection

benchmark needs to be made pragmatically and under the consideration that it may be inaccurate for some target variables.

6.2.5 Alternative approaches to using measurement and selection benchmarks

Survey methodology has made available alternative approaches to evaluating survey error components. As far as selection error against the sampling frame is concerned, two alternative approaches were reviewed and applied in chapter 2: using auxiliary data and extrapolating benchmark measurements to the sampling frame. Auxiliary data needs to be ideally available on sampling frame level which often restricts analyses to socio-demographic and regional characteristics. Whereas selection analyses on socio-demographics are standard in survey methodology, such analyses may not be informative about selection error on target variables, unfortunately, especially if correlations between both types of variable are low. A second approach applies available auxiliary data to extrapolate information of late respondents or nonrespondents to the population to estimate selection error (Billiet, Philippens, Fitzgerald, & Stoop, 2007; Groves, 2006; Peress, 2010). These approaches help to assess absolute selection error against the population true score, when the selection benchmark alone is viewed as insufficient.

As an alternative to choosing a measurement benchmark, so-called data quality indicators (DQI) have been used as indicators for the size of measurement error in the absence of true scores. Typical DQIs are response styles, such as acquiescent responding, extreme responding, mid-point responding, straight lining, or item omission (item nonresponse). The advantage of this line of research may be that an estimate of the 'absolute size of measurement error' is available (e.g., a measure for the degree of 'acquiescence' in a survey mode), but under this perspective 'measurement errors' receive a fundamentally different interpretation than in the TSE framework. However, it has been shown that response styles may crucially affect the extent of systematic measurement error across sets of questions. Chapter 3 illustrated one method for estimating such effects, discussing measurement effects on systematic erro across sets of questions partly as a result of differences in DQI across modes. The evaluation of DQIs can strongly support analyses of measurement error in the absence of clear benchmarks and may also represent another criterion for choosing measurement

benchmarks for the analysis of TSE (i.e., modes scoring well on DQIs may represent proper measurement benchmarks).

As discussed above, setting a measurement benchmark is particularly important for giving mode differences in systematic error (bias) a substantial interpretation. When error variance (e.g., variable error of a mean estimator due to measurement error) is concerned, other models are available and setting a measurement benchmark is not required. One approach uses latent variable models that allow estimating method- and item-level variance components, such as the multiple-group confirmatory factor analysis (MCFA) model applied in chapter 3 or multi-trait-multi-method (MTMM) models. Originating in psychological literature, estimates based on random and systematic error variance are called validity and reliability parameters (Alwin, 2007; Saris & Andrews, 1991). The advantage of these indicators is that they allow absolute conclusions about mode effects on measurement error variance, where generally modes with smaller random and systematic measurement error variance are to be preferred (i.e., representing the more reliable and valid measurement instruments). However, a disadvantage of this perspective is the negligence of measurement bias components due to the sole focus on disentanglement of true score, systematic, and random measurement error variance.

6.3 Generalizing the Re-interview Approach for Decomposing Mode Effects

In this dissertation, a re-interview approach (or within-subject design approach) for decomposing measurement and selection effects between single-mode surveys and in sequential mixed-mode designs was developed (chapters 2, 4, and 5). Re-interview data can be treated as record check data with missing information due to nonresponse (chapter 2) or used as auxiliary information exploited in the estimation of single- and mixed-mode measurement and selection effects (chapters 4 and 5).

The MEPS experiment set the context for a relatively large number of single- and mixed-mode comparisons made throughout chapters 2 to 5. However, the types of mixed-mode designs that could be evaluated were quite specific (i.e., sequential mixed-mode, where web, mail, and telephone were followed-up by face-to-face, respectively;

and effects due to single-to-single mode switches). It is now discussed, whether the re-interview approach can be useful for evaluating survey error in other and more complex types of mixed-mode designs (in analogy to Figure 6.1).

An important factor in the generalization of the re-interview approach is practical feasibility. Since, in practice any re-interview survey might be expensive (e.g., face-to-face), a full re-interview of all units, as in the MEPS experiment, is not cost efficient. This holds in particular, if a decomposition of effects needs to be executed for many surveys or repeatedly over time, and not only as a singular experiment at a given point in time. For this reason, it may be an option to re-interview only a sub-sample of respondents in a mixed-mode design (Figure 6.2). Figure 6.2 demonstrates the missing data pattern of the sequential and concurrent designs illustrated in Figure 6.1. Contrary to the situation discussed in chapters 4 and 5, not all respondents in modes A or B are re-interviewed. Thus some outcomes remain 'potential' (grey areas on Y^C). It is also clear that the designs can both involve a re-interview independent from the total number of modes used in the designs. Mixed-mode designs with more or less modes than in this example can therefore involve similar re-interview components.

As discussed in section 6.2, mode effects are defined either against a single-mode benchmark or against a measurement benchmark within the mixed-mode design. If a single-mode benchmark is available, the goal is to estimate potential outcomes of single-mode Y^C for respondents in the mixed-mode design. This procedure requires additionally a re-interview of a sub-sample of respondents in the single-mode benchmark C (not shown in Figure 6.2). The details of estimation are addressed in chapter 5 for a sequential design with two modes.

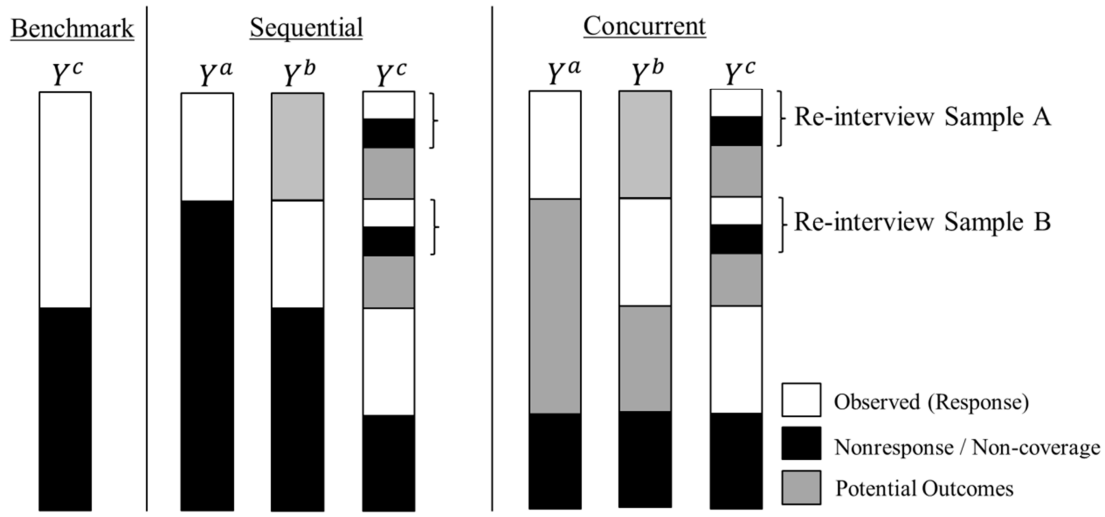


Figure 6.2: Examples of a re-interview in sub-samples of respondents to sequential and concurrent mixed-mode surveys (with an optional single-mode benchmark survey)

If a single-mode benchmark is not available, a mode in the sequential and concurrent designs is defined as measurement benchmark (e.g., mode C). The re-interview data may then be used in two different ways. First, the repeated measures may be applied as benchmark data themselves, if exact replications of target variables can be realized (Y^C). Missing data on Y^C due to nonresponse in the re-interview and due to the sub-sampling can be adjusted using the 'forward method' described in chapter 4, leading to a complete data vector Y^C for respondents in the mixed-mode designs. Since this data is assumed to be the measurement benchmark, relative measurement and selection effects can be estimated. Essentially, this approach creates 'records' to be used as in a record-check approach with administrative information. A disadvantage of the re-interview benchmark approach is the need for measurement equivalence and the need to repeat all relevant target-questions in the re-interview, so that the re-interview may become long suggesting high response burden for repeated respondents (cf. chapter 4).

In addition, the repeated interview in mode C may be regarded the selection benchmark (e.g., a face-to-face survey). Instead of the independent single-mode survey, which is not available, now the re-interview represents the selection benchmark. This step may require careful fieldwork design to allow an assumption of response independence in mode C from the earlier response in the mixed-mode design.

An alternative approach to using re-interview data as benchmark is using it only as auxiliary information applied for estimating potential outcomes under modes A and B (cf. chapter 5). This approach is based on less strict assumptions (e.g., no measurement or response equivalence / independence from the earlier waves). The objective then becomes to build measurement models of Y^A and Y^B on the observed part of the data from the re-interview and extrapolate this information to the unobserved outcomes under both modes (e.g., predict or impute). Based on completed data vectors Y^A and Y^B, relative measurement and selection effects are estimated against the measurement benchmark Y^C of respondents in the mixed-mode design (i.e., not using Y^C from the re-interview, but only Y^C from the sub-set of respondents in mode C in the mixed-mode design).

An open question is, which types of variables should be collected in the re-interview for this purpose. Three types of variables may be particularly fruitful. First, approximate re-measurement of Y^C is possible in the re-interview, where it is acknowledged that an exact measurement of Y^C may not be possible (e.g., if measurement equivalence between the mixed-mode follow-up in mode C and the re-interview in mode C does not hold exactly). Still, approximate Y^C may be strong predictor variables in the sense that conditioning on approximate Y^C allows assuming MAR Y^A and Y^B in the response set of the sequential or concurrent designs. A more general type of data allowing this assumption are called backdoor variables in the causal inference literature. Contrary to Y^C, backdoor variables include any other information that may explain the relative selection mechanism between modes (e.g., why a respondent did not respond in mode A or B but replied in mode C). A final candidate variable type is called 'front-door variables' (Vannieuwenhuyze et al., 2014). Front-door variables have not been addressed in the present research. Instead of explaining a selection mechanism (backdoor variables), front-door variables explain why measurement effects occur (i.e., they represent the causal mechanism behind occurrence of measurement effects). Their use in the missing data situation illustrated in Figure 6.2 is not fully explored yet, however.

Regardless of perspectives, any sampling frame information (referred to as X in chapters 4 and 5) may be applied to additionally extrapolate mode effect estimates to the

population (i.e., adjust for unit nonresponse). In particular, this step may allow estimating absolute selection error against the sampling frame (based on the assumption of missing at random [MAR] data conditional *X*). However, doing so requires further careful considerations about the underlying MAR assumptions on available sampling frame information.

The new types of re-interview designs suggested raise several questions relevant for further research. The most central questions are discussed in section 6.5.

6.4 Inferential Fallacies about Measurement Effects in Split-Ballot Designs

The split-ballot design, also called between-subject design, takes on a central role in methodological research about measurement effects of survey modes (chapter 4). Experimental mode comparisons generally randomize subjects into modes in order to draw inferences about measurement effects and mode differences in data quality indicators (DQIs). In addition to simple comparisons between modes, methodologists often try to draw inferences about the likely effects to be expected in mixed-mode designs based on the estimates of measurement effects obtained in split-ballot mode comparison designs. Also the MEPS experiment represented a split-ballot design with four conditions and it was one objective of the experiment to optimize the design of the Crime Victimization Survey based on measurement effect estimates (cf. section 1.5.1). The centrality of the split-ballot design in mixed-mode research and the MEPS experiment warrants a retrospective discussion of the type of inferences that are possible from split-ballot mode experiments to measurement effects in mixed-mode designs and potential inferential fallacies in these respects.

In principle, two types of split-ballot designs can be distinguished: the field and the lab split-ballot design (Figure 6.3). The primary purpose of split ballot designs is causal inference about treatment effects and, in the situation of mode experiments, inference about measurement effects. A strong advantage of lab experiments is full control over the randomization process. If full randomization of subjects to modes is possible, unconfoundedness of assignment mechanism and target variables always applies. Therefore, unbiased estimation of measurement effects is possible. However, lab

experiments, usually recruit relatively small numbers of individuals in non-random ways from often very selective groups (e.g., students) and randomize into modes only after successful recruitment. Field experiments, to the contrary, administer modes as in other general surveys. Due to the selective mechanisms of modes discussed in this book, full control over randomization is not possible in fielded mode experiments. Measurement effects are still confounded with selection effects (cf. chapter 4).

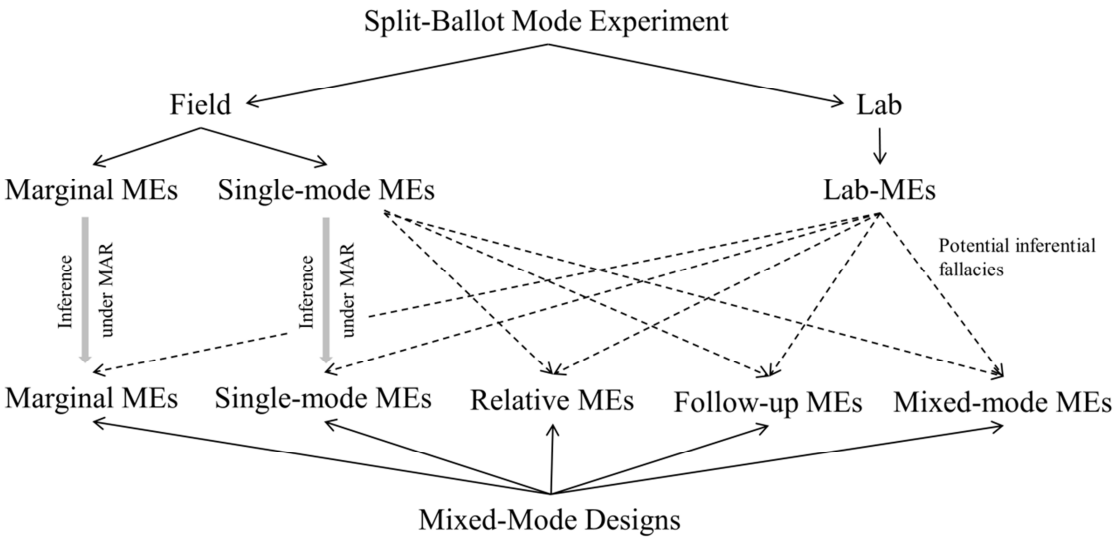


Figure 6.3: Illustration of potential inferential fallacies about measurement effects (MEs) in mixed-mode designs based on split-ballot mode experiments (dashed arrows)

It is now instructive to consider the type of measurement effects estimated in field and lab split-ballot experiments and compare these effects on theoretical level to measurement effects encountered in mixed-mode designs (section 6.2). As discussed in chapter 4, split-ballot field experiments (Figure 6.3, left) may estimate marginal measurement effects based on a missing at random (MAR) assumption that allows extrapolating observed sample estimators to the sampling frame (e.g., using sampling frame auxiliary information). Marginal measurement effects can be defined as the unconditional difference of mode-specific response distributions in the population. For example, in chapter 3, propensity score weighting of the MEPS data was used for drawing inference about population-level measurement effects on systematic and random measurement error. Chapter 4 defined marginal measurement effects as unconditional difference in the expectation of mode-specific response distributions. A theoretically strong point of marginal measurement effects is that after successful

extrapolation of response sample estimators to the population, it is irrelevant[1], from which design mode-specific responses were taken to estimate marginal measurement effects. Under the assumption that the population calibration model is correct (e.g., in terms of MAR assumptions) in, both, a split-ballot and a mixed-mode design, marginal measurement effect estimates are identical in expectation. Marginal effects from a split-ballot design could thus form the basis of speculation about effects in mixed-mode surveys after calibration to the sampling frame. However, the assumption that the calibration model in both types of designs is correct, can often not be held in practice, so that evaluating marginal measurement effects in split ballot designs may not be feasible or plausible in practice (cf. chapter 4).

A solution is offered by focusing on conditional measurement effects instead. Chapters 4 and 5 introduced assumptions for unbiased estimation of single-mode measurement effects. Besides estimation under plausible MAR assumptions in re-interview designs, another advantage of single-mode measurement effects is that the effects feature a direct interpretation in sequential mixed-mode designs. For example, if 'face-to-face' is the single-mode benchmark, a single-mode measurement effect against 'web' indicates the extent of measurement error bias introduced by starting a sequential mixed-mode survey by web instead of using face-to-face directly (the measurement effect is estimated conditional on web response; cf. chapter 4 and 5). As such, single-mode effects quantify the degree of total bias, measurement error bias, and selection error bias that needs to be prevented or compensated by appropriate mixed-mode questionnaire design (cf. section 6.2.3). An advantage of fielded mode experiments is that these effects may be estimated. In addition, split ballot designs are a natural way to evaluate single-to-single mode switches in repeated cross-sectional or longitudinal survey, which also require the evaluation of single-mode effects (cf. Figure 6.1 and discussion in section 6.2.1).

[1] It should be specified that this statement assumes that the mode-specific population distributions of potential outcomes in a mixed-mode design are equivalent to the distribution in the split-ballot experiment. If measurement error is strongly changed in conditions of the mixed-mode design relative to the split ballot experiment, this statement does not hold. New and unexplored designs as well as scale influences, because of smaller samples and interviewers that need to become acquainted to different circumstances, require careful design of field experiments to guarantee comparability.

However, as illustrated in Figure 6.3, care needs to be taken when drawing conclusions about relative, follow-up or even mixed-mode measurement effects from conditional measurement effect estimates in split-ballot field experiments. The exact selection mechanisms in mixed-mode designs may be very different from single-mode selection mechanisms. For this reason, split ballot experiments cannot fully predict the expected relative measurement effects between modes in a sequential or concurrent mixed-mode design. In other words, exact planning of mixed-mode surveys is not fully possible by split-ballot mode experiments. This step always requires an implementation of different candidate mixed-mode designs in the end.

Still, it should be said that measurement equivalence between candidate modes and a measurement benchmark mode in split-ballot designs encourages the combination of these modes in mixed-mode data collection. In technical terms, however, the relative, follow-up, and mixed-mode measurement effects may turn out to be very different in sequential and concurrent mixed-mode designs than estimates of single-mode measurement effects in split-ballot field experiments may suggest.

The possibility for inferential fallacies about measurement effects has to be considered even more strongly in lab than in fielded mode experiments (Figure 6.3, right). Lab experiments involve strong self-selection of subjects to participation and most often the selection mechanism is simply unknown. Extrapolating effect estimators from lab experiments to the population is practically impossible, and inference to single-mode, relative or any other effects may likewise be flawed. It should be noted, however, that lab experiments are based on the assumption that the underlying causal model of measurement effects works essentially equivalent in the population and all sub-domains of interest, so that experimentation with selective sub-groups in labs does not preclude conclusions about measurement effects. Experiments in ‘questionnaire labs’ represent an important aspect of mixed-mode research for this reason. If the causal mechanism behind measurement effects is understood the prevention of measurement effects may be significantly facilitated regardless of the eventual mixed-mode design used in the field. An additional advantage of lab experimentation is that measurement effects may be more easily and cost-efficiently studied than in fielded mode experiments or mixed-mode surveys.

Given the centrality of mode experiments in mixed-mode research, the reviewed fallacies in field and lab split-ballot experiments should receive more general awareness among survey researchers to avoid false conclusions about measurement effect estimates. Lab and field split-ballot experiments should be regarded as a first step towards evaluating measurement effects and obtaining equivalence in mixed-mode designs. However, the ultimate step in designing mixed-mode surveys needs to represents an experimental implementation of different candidate mixed-mode strategies, because only when the correct selection mechanisms and effects of the final mode-specific circumstances are observed in a mixed-mode design can final conclusions about the size of measurement effects be drawn.

6.5 Outlook

The previous sections reviewed the central contributions of the present dissertation to the field of mixed-mode research and discussed them in the context of general mixed-mode design problems. This discussion raises important new questions for further empirical research and methodological development. This section summarizes and discusses the most central lines for future research.

6.5.1 Need for empirical replications

Section 6.1 reviewed the empirical contributions of the present dissertation, which clearly suggested that measurement effects were the dominant problem in the Crime Victimization Survey. It was suggested to interpret these results with care, because measurement and selection effects may be very different in surveys with other topics. For this reason, there is need for empirical replications of the empirical findings on other survey topics and populations.

However, any replication study does not necessarily need to follow the design of the MEPS experiment. Rather, the above discussion suggests that alternative re-interview designs may be effective in decomposing measurement and selection effects in mixed-mode surveys. The focus of the MEPS experiment originally was a split-ballot (between-subject) comparison with re-interview follow-up. The discussion in section 6.4 has made clear that it is important to observe mode effects also in the eventual

candidate mixed-mode designs. Therefore, future research could also focus on estimating measurement and selection effects in mixed-mode surveys and not necessarily use split-ballot mode comparisons.

6.5.2 Choosing benchmark modes based on empirical indicators

Section 6.2 advocated the use of measurement and selection benchmarks in the evaluation of mode effects. The choice of a measurement benchmark mode is crucial in the absence of true scores to give mode effects a more direct interpretation, but it often involves a more or less pragmatic decision on a preferred measurement or trusted selection mode. Nevertheless, it would be desirable to come to more objective conclusions about the ideal benchmark mode than the ones provided so far. In section 6.2.4, it was argued, for example, that data quality indicators (DQIs) may support making objective decisions about measurement benchmarks. One outcome of research about mode-specific differences in response styles may represent indicators that act as selection criteria for measurement benchmark modes. More generally, there is need for both conceptual and empirical approaches to help practitioners choose benchmarks.

6.5.3 Adjusting measurement effects towards measurement benchmarks

Measurement effects in mixed-mode surveys imply an increase in measurement error relative to the measurement benchmark. This increase may outweigh any gains made by selection effects (regardless of whether evaluated as relative selection effects or against a selection benchmark). As argued, the first step in mixed-mode questionnaire design is preventing the occurrence of measurement effects by minimizing systematic measurement error at the level of the measurement benchmark mode. Nevertheless, prevention of measurement effects may not be feasible, especially for the case of interviewer and self-administered modes that showed very consistent occurrence of measurement effects in the present study. In practice, it might not be possible to fully avoid all relative measurement effects between these mode groups, for example. For this situation it is desirable to be able to adjust measurement effects at the estimation stage of the survey process towards the measurement benchmark mode (e.g., when estimating population means, totals, or covariance-based statistics). For example, in a mixed-mode

design using web, telephone, and face-to-face, one goal of the adjustment step may be to create an estimator of face-to-face responses from the mixed-mode response set, thus adjusting answers under web and telephone to face-to-face.

The statistical methods for adjusting measurement effects represent a field for further research. However, the frameworks discussed in the present dissertation suggest that prediction or imputation of potential outcomes in mixed-mode surveys is one possible approach towards adjustment of measurement effects (cf. Figure 6.2). Similar approaches have recently been suggested by Suzer Gurtekin (2013) and Kolenikov and Kennedy (2014). In addition, Bayesian approaches may be applied to include the uncertainty of predictions and imputations of potential outcomes in final estimates. In multiple imputation approaches, for example, such uncertainty in imputed potential outcomes is reflected by the between imputation variance across multiply imputed data sets (cf. chapter 5). Nevertheless, other approaches for adjusting measurement effects may be developed in the future.

An essential step in the adjustment of measurement effects is that the data in a mixed-mode design can be considered MAR conditional on the observed exogenous auxiliary information. In other words, it must be possible to model measurement effects based on observed outcomes and this measurement model also has to apply in the missing part of the data. This step requires the availability of plausible adjustment data. The re-interview approach is one important way for making available such data (section 6.3). Therefore, this approach may not only turn out to be fruitful in the estimation of mode effects in mixed-mode surveys, but also for the adjustment of measurement effects in ongoing mixed-mode field research.

6.5.4 Using re-interview data for estimating and adjusting measurement effects

Re-interview data have potential to deliver important auxiliary information for estimating and adjusting measurement effects in ongoing mixed-mode surveys. Section 6.3 discussed how a re-interview design could be used to sub-sample a part of respondents in a mixed-mode design for a second survey (cf. Figure 6.2). This idea should receive further scholarly investigations. Several research questions appear relevant in both respects.

First, it is essential to evaluate the interplay of the strength of the correlation between survey variables and re-interview variables, sample size of the re-interview sample, and efficiency (e.g., standard errors) of effect estimates and adjusted estimators. Put differently, it has to be known how strong measurement effect models must be to come to efficient imputations (or predictions) in the mixed-mode design given the size of the re-interview sample and potentially attenuated correlations between repeated measurement occasions due to random error. A related aspect is costs, which directly are connected to the size of the re-interview sample. On a basic level, it is to be expected that adjustments are more efficient (e.g., smaller standard error) for larger re-interview samples, which, however, raises costs. It would be desirable to find an optimal balance between these constraints.

Second, the type of models or estimators used for measurement effects and for adjusting statistics should be evaluated. One question is, for example, if imputation is an efficient adjustment method or if other methods may be preferable.

Third, section 6.3 discussed different types of variables (re-measurements, backdoor variables, front-door variables) that may be used in adjustment models. The exact type of variables (or questions) to be included in the re-interview questionnaire is, however, not fully clear. An important constraint in this respect may be the length of the re-interview, which should be short to keep response burden for repeated respondents low.

Finally, practical considerations should be made. For example, it should be evaluated, whether it is feasible in survey practice to administer re-interviews next to ongoing mixed-mode surveys and which organizational and legal constraints of including a re-interview in a mixed-mode design would need to be taken into account.

6.6 Concluding Remarks

The methodology of mixed-mode surveys is a relatively new and unexplored field of research. Even though the first reports on the use of mixed-mode data collection designs date back to the 1960s and 70s (Hochstim, 1967; Siemiatycki, 1979), the relevance of the role of the survey mode in a survey design has only recently received wider attention among methodologists, fostered mainly by practical developments, such as the increasing availability of modes, technological facilitation of mixed-mode data

collection, and the general need to exchange the more traditional and costly modes by inexpensive data collection strategies. Especially in the first decade of the 21st century, world-wide survey research has seen a rapid increase in the use of mixed-mode designs following primarily pragmatic decisions taken by survey practitioners.

Survey methodology has tried to keep up with this rapid development, but the immense need for answers to methodological problems could not be matched by systematic methodological developments. At the outset of this research project, the field of mixed-mode research could be characterized as very fragmented into different sub-streams. Much methodological work focussed on singular, ad-hoc answers to research problems clustering relatively vaguely around measurement equivalence and representativeness of different types of mixed-mode surveys. The problem of confounded measurement and selection effects had received some attention, but no formal definition, and the problem often remained unaddressed in empirical applications. Finally, the situation was and still is strongly characterized by the assumption that 'mode effects' observed in case studies, often based on split-ballot designs, may be strongly generalizable to other populations, survey topics, and questionnaires, and that such experiences may have general implications for mixed-mode survey design.

I conclude this dissertation with advocating a more systematic approach to the problem of mixed-mode design. I suggest that methodological development needs to take a more central role in mixed-mode research than singular empirical case studies with potentially weak generalizability and relatively unsystematic use of methodology. Such development should primarily focus around two larger lines of research.

First, it is essential to develop and do empirical research in a more systematically defined statistical framework. This dissertation and further research on mode effects of the recent years have made progress on theoretical approaches to studying survey error in mixed-mode designs (Schouten et al., 2013; Vannieuwenhuyze & Loosveldt, 2013). Nevertheless, this development has not come to an end. There is still great need to integrate different perspectives on survey error into a more general framework under which systematic research into mode effects may be conducted more efficiently in the future. Such a framework needs to take into account mode-specific selection

mechanisms in different mixed-mode designs, different causes to selection (e.g., coverage, contact, and refusal), the role of true score and psychometric models, the role of benchmarks (measurement or selection), the type of estimator (e.g., mean, variance, correlations), and a clear differentiation of the estimator's systematic and random error components. Finally, related perspectives on data quality indicators, such as response styles, may need to be integrated or at least find a place and systematic interpretation in a more general mixed-mode research context.

Second, further approaches need to be developed to support the design of mixed-mode surveys prior to implementation as well as for monitoring ongoing mixed-mode surveys. The problem of measurement effects will continue to represent a key problem under this perspective. In the design of mixed-mode questionnaires, measurement effects may be prevented. Such efforts may start in laboratory situations, but ultimately their success has to be evaluated in the field. If prevention turns out impossible, another option is to fully avoid problematic modes or to adjust effects after they have occurred. The coping with measurement effects in mixed-mode surveys could receive attention under the more general heading of 'estimation from mixed-mode data'. Estimation can be discussed under the perspective of optimal unit nonresponse adjustment in mixed-mode surveys as well as optimal adjustment of measurement effects against measurement benchmarks. An ideal protocol for estimation from mixed-mode survey data would integrate methods for the adjustment of measurement effects with methods for extrapolating response sample estimates to the reference population. Such methodology may be developed in the future.

An advantage of a stronger methodological perspective on mixed-mode surveys is that singular and empirically motivated advice on mixed-mode design can be supplemented by standard sets of methods to be used in survey practice for particular survey topics and populations. Nevertheless, these suggestions should not imply that empirical research on mixed-mode surveys should be fully replaced by purely theoretical and methodological developments. Empirical case studies may still lead to best practice advice on mixed-mode design, which is strongly needed when taking design decisions. Two important ways to generalize findings from empirical research are meta-analyses and systematic literature reviews of effects. However, such analyses strongly require comparability of error estimators and quality indicators. It is, therefore,

important for empirical research to follow a common framework and research agenda in order to allow comparability and generalizability of results across case studies and populations.

Given the rapid development of technology, it is likely that mixed-mode data collection will receive a fixed place in the toolkit of survey researchers. The currently quickly diversifying web mode to surveys on many different types of electronic devices, such as mobile telephones and tablet PCs, which will probably have to be used concurrently in the surveys of the future, is only one example of the continued importance of the mixed-mode topic for further methodological research. It is likely that the next ten years of survey research will proceed as ground breaking as the past decades. Methodologists need to accept the challenge to evaluate, maintain, and improve the quality of survey data collection under these constantly changing circumstances.

References

Adua, L., & Sharp, J. S. (2010). Examining survey participation and response quality: The significance of topic salience and incentives. *Survey Methodology*, *36*(1), 95–109.

Agresti, A. (2002). *Categorical Data Analysis* (2nd ed.). Hoboken, New Jersey: Wiley-Interscience.

Alwin, D. F. (2007). *Margins of Error*. Hoboken: Wiley.

Andrews, F. M. (1984). Construct Validity and Error Components of Survey Measures: A Structural Modeling Approach. *Public Opinion Quarterly*, *48*(2), 409–442. doi:10.1086/268840

Aquilino, W. S. (1994). Interview Mode Effects in Surveys of Drug and Alcohol Use: A Field Experiment. *Public Opinion Quarterly*, *58*(2), 210–240. doi:10.1086/269419

Asparouhov, T. (2005). Sampling Weights in Latent Variable Modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *12*(3), 411–434. doi:10.1207/s15328007sem1203_4

Asparouhov, T., & Muthén, B. O. (2006). *Robust Chi Square Difference Testing with Mean and Variance Adjusted Test Statistics*. Mplus Web Notes. Retrieved from http://www.statmodel.com/download/webnotes/webnote10.pdf

Bakker, B. F. M. (2012). Estimating the validity of administrative variables. *Statistica Neerlandica*, *66*(1), 8–17. doi:10.1111/j.1467-9574.2011.00504.x

Bälter, K. A., Bälter, O., Fondell, E., & Lagerros, Y. T. (2005). Web-based and Mailed Questionnaires. *Epidemiology*, *16*(4), 577–579. doi:10.1097/01.ede.0000164553.16591.4b

Berrens, R. P., Bohara, A. K., Jenkins Smith, H., Silva, C., & Weimer, D. L. (2003). The Advent of Internet Surveys for Political Research: A Comparison of Telephone and Internet Samples. *Political Analysis*, *11*(1), 1 –22. doi:10.1093/pan/11.1.1

Bethlehem, J. (1988). Reduction of Nonresponse Bias Through Regression Estimation. *Journal of Official Statistics*, *4*(3), 251–260.

Bethlehem, J. (2002). Weighting Nonresponse Adjustment Based on Auxiliary Information. In R. M. Groves, D. A. Dillman, J. Eltinge, & R. J. A. Little (Eds.), *Survey Nonresponse*. New York: Wiley & Sons.

Bethlehem, J., & Biffignandi, S. (2011). *Handbook of Web Surveys*. John Wiley & Sons, Inc.

Bethlehem, J., Cobben, F., & Schouten, B. (2011). *Handbook of Nonresponse in Household Surveys*. New Jersey: Wiley.

Biemer, P. P. (1988). Measuring Data Quality. In R. M. Groves, P. P. Biemer, L. Lyberg, J. T. Massey, W. Nicholls, II, & J. Waksberg (Eds.), *Telephone Survey Methodology* (pp. 341–375). New York: Wiley.

Biemer, P. 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.

Biemer, P. P. (2010a). Overview on Design Issues: Total Survey Error. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of Survey Research* (pp. 27–58). Emerald.

Biemer, P. P. (2010b). Total Survey Error: Design, Implementation, and Evaluation. *Public Opinion Quarterly*, *74*(5), 817–848. doi:10.1093/poq/nfq058

Biemer, P. P., & Lyberg, L. E. (2003). *Introduction to Survey Quality*. New Jersey: Wiley.

Biemer, P. P., & Stokes, L. (1991). Approaches to the Modeling of Measurement Errors. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement Errors in Surveys* (pp. 487–517). Hoboken, New Jersey: Wiley.

Billiet, J. B., & Davidov, E. (2008). Testing the Stability of an Acquiescence Style Factor Behind Two Interrelated Substantive Variables in a Panel Design.

Sociological Methods & Research, *36*(4), 542–562. doi:10.1177/0049124107313901

Billiet, J. B., & McClendon, M. J. (2000). Modeling Acquiescence in Measurement Models for Two Balanced Sets of Items. *Structural Equation Modeling: A Multidisciplinary Journal*, *7*(4), 608–628. doi:10.1207/S15328007SEM0704_5

Billiet, J. B., Philippens, M., Fitzgerald, R., & Stoop, I. (2007). Estimation of nonresponse bias in the European Social Survey: using information from reluctant respondents. *Journal of Official Statistics*, *23*(2), 135–162.

Blalock, H. M. (1970). A Causal Approach to Nonrandom Measurement Errors. *The American Political Science Review*, *64*(4), 1099–1111. doi:10.2307/1958360

Blumberg, S. J., Luke, J. V., Cynamon, M. L., & Frankel, M. R. (2008). Recent Trends in Household Telephone Coverage in the United States. In J. Lepkowski, C. Tucker, J. M. Brick, E. De Leeuw, L. Japec, P. J. Lavrakas, … R. L. Sangster (Eds.), *Advances in Telephone Survey Methodology* (pp. 56–86). New York: Wiley.

Blyth, B. (2008). Mixed mode: the only "fitness" regime. *International Journal of Market Research*, *50*(2), 241–266.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York: Wiley.

Bollen, K. A., & Paxton, P. (1998). Detection and Determinants of Bias in Subjective Measures. *American Sociological Review*, *63*(3), 465–478. doi:10.2307/2657559

Bowling, A. (2005). Mode of questionnaire administration can have serious effects on data quality. *Journal of Public Health*, *27*(3), 281 –291. doi:10.1093/pubmed/fdi031

Braunsberger, K., Wybenga, H., & Gates, R. (2007). A comparison of reliability between telephone and web-based surveys. *Journal of Business Research*, *60*(7), 758–764. doi:10.1016/j.jbusres.2007.02.015

Buchanan, T., Johnson, J. A., & Goldberg, L. R. (2005). Implementing a Five-Factor Personality Inventory for Use on the Internet. *European Journal of Psychological Assessment*, *21*(2), 115–127. doi:10.1027/1015-5759.21.2.115

Buchanan, T., & Smith, J. L. (1999). Using the Internet for psychological research: Personality testing on the World Wide Web. *British Journal of Psychology*, *90*(1), 125–144. doi:10.1348/000712699161189

Buelens, B., & Brakel, J. A. van den. (2014). Measurement Error Calibration in Mixed-mode Sample Surveys. *Sociological Methods & Research*. doi:10.1177/0049124114532444

Buelens, B., Van der Laan, J., Schouten, B., Van den Brakel, J., & Klausch, T. (2012). *Disentangling mode-specific selection and measurement bias in social surveys* (Discussion paper No. 201211). The Hague, The Netherlands: Statistics Netherlands. Retrieved from http://cbs.nl/NR/rdonlyres/987B3CE9-1FB0-44B3-B4DE-D91C8ADC9794/0/201211x10pub.pdf

Busse, B., & Fuchs, M. (2011). The components of landline telephone survey coverage bias. The relative importance of no-phone and mobile-only populations. *Quality & Quantity*, *46*(4), 1209–1225. doi:10.1007/s11135-011-9431-3

Chang, L., & Krosnick, J. A. (2009). National Surveys Via Rdd Telephone Interviewing Versus the Internet. *Public Opinion Quarterly*, *73*(4), 641–678. doi:10.1093/poq/nfp075

Chen, F. F. (2007). Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464–504. doi:10.1080/10705510701301834

Christian, L. M., Dillman, D. A., & Smyth, J. D. (2008). The Effects of Mode and Format on Answers to Scalar Questions in Telephone and Web Surveys. In J. Lepkowski, C. Tucker, J. M. Brick, E. D. De Leeuw, L. Japec, P. J. Lavrakas, … R. L. Sangster (Eds.), *Advances in Telephone Survey Methodology* (pp. 250–275). New York: Wiley.

Cochran, W. G. (1977). *Sampling Techniques* (2nd ed.). New York: Wiley.

Cohen, C. (1977). *Statistical Power Analysis For The Behvioral Sciences*. New York: Academic Press.

Cole, M. S., Bedeian, A. G., & Feild, H. S. (2006). The Measurement Equivalence of Web-Based and Paper-and-Pencil Measures of Transformational Leadership. *Organizational Research Methods*, *9*(3), 339–368. doi:10.1177/1094428106287434

Couper, M. P. (2000). Review: Web Surveys: A Review of Issues and Approaches. *Public Opinion Quarterly*, *64*(4), 464–494. doi:10.1086/318641

Couper, M. P., Kapteyn, A., Schonlau, M., & Winter, J. (2007). Noncoverage and nonresponse in an Internet survey. *Social Science Research*, *36*(1), 131–148. doi:10.1016/j.ssresearch.2005.10.002

Couper, M. P., & Miller, P. V. (2008). Web Survey Methods Introduction. *Public Opinion Quarterly*, *72*(5), 831–835. doi:10.1093/poq/nfn066

Davis, D. W. (1997). Nonrandom Measurement Error and Race of Interviewer Effects Among African Americans. *The Public Opinion Quarterly*, *61*(1), 183–207. doi:10.1086/297792

De Beuckelaer, A., & Lievens, F. (2009). Measurement Equivalence of Paper-and-Pencil and Internet Organisational Surveys: A Large Scale Examination in 16 Countries. *Applied Psychology*, *58*(2), 336–361. doi:10.1111/j.1464-0597.2008.00350.x

De Leeuw, E. (1992). *Data Quality in Mail, Telephone, and Face to Face surveys*. Amsterdam: TT-Publicaties.

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. (2008). Choosing the method of data collection. In E. De Leeuw, J. J. Hox, & D. A. Dillman (Eds.), *International Handbook of Survey Methodology* (pp. 113–135). New York: Taylor & Francis.

De Leeuw, E., & De Heer, W. (2002). Trends in Household Survey Nonresponse: A Longitudinal and International Comparison. In R. M. Groves, D. A. Dillman, J. Eltinge, & R. Little (Eds.), *Survey Nonresponse* (pp. 41–54). New York: Wiley.

De Leeuw, E., Dillman, D. A., & Hox, J. J. (2008). Mixed mode surveys: When and Why. In E. De Leeuw, D. A. Dillman, & J. J. Hox (Eds.), *International Handbook of Survey Methodology* (pp. 299–316). New York: Lawrence Erlbaum.

De Leeuw, E., & Hox, J. (2008). Self Administered Questionnaires: Mail Surveys and Other Applications. In E. De Leeuw, D. A. Dillman, & J. J. Hox (Eds.), *International Handbook of Survey Methodology* (pp. 299–316). New York: Lawrence Erlbaum.

De Leeuw, E., & Hox, J. J. (2011). Internet Surveys as Part of a Mixed-Mode Design. In M. Das, P. Ester, & L. Kaczmirek (Eds.), *Social and Behavioral Research and the Internet: Advances in Applied Methods and Research Strategies* (pp. 45–76). New York: Routledge.

De Leeuw, E., Hox, J. J., & Dillman, D. A. (Eds.). (2008). *The International Handbook of Survey Methodology*. New York: Taylor & Francis.

De Leeuw, E., Hox, J. J., & Scherpenzeel, A. C. (2011). Emulating Interviewers in an Online Survey: Experimental Manipulation of "Do-Not-Know" over the Phone and on the Web. In American Statistical Association (Ed.), *JSM Proceedings, Survey Research Methods Section* (pp. 6305–6314). Alexandria, VA: American Statistical Association.

De Leeuw, E., Mellenbergh, G. J., & Hox, J. J. (1996). The Influence of Data Collection Method on Structural Models. *Sociological Methods & Research*, *24*(4), 443 – 472. doi:10.1177/0049124196024004002

Deutskens, E., de Ruyter, K., & Wetzels, M. (2006). An Assessment of Equivalence Between Online and Mail Surveys in Service Research. *Journal of Service Research*, *8*(4), 346 –355. doi:10.1177/1094670506286323

Dijkstra, W., & Smit, J. H. (2002). Persuading Reluctant Recipients in Telephone Surveys. In R. M. Groves, D. A. Dillman, J. Eltinge, & R. Little (Eds.), *Survey Nonresponse*. New York: Wiles & Sons.

Dillman, D. A. (1991). The Design and Administration of Mail Surveys. *Annual Review of Sociology*, *17*, 225–249.

Dillman, D. A. (2009). Some Consequences of Survey Mode Changes in Longitudinal Surveys. In P. Lynn (Ed.), *Methodology of Longitudinal Surveys* (pp. 127–140). Chichester: Wiley.

Dillman, D. A., & Christian, L. M. (2005). Survey Mode as a Source of Instability in Responses across Surveys. *Field Methods*, *17*(1), 30–52. doi:10.1177/1525822X04269550

Dillman, D. A., & Messer, B. L. (2010). Mixed-Mode Surveys. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of Survey Research* (pp. 27–58). Emerald.

Dillman, D. A., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J., & Messer, B. L. (2009). Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the Internet. *Social Science Research*, *38*(1), 1–18. doi:10.1016/j.ssresearch.2008.03.007

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2009). *Internet, Mail, and Mixed-Mode Surveys. The Tailored Design Method.* New Jersey: Wiles & Sons.

Dillman, D. A., West, K. K., & Clark, J. R. (1994). Influence of an Invitation to Answer by Telephone on Response to Census Questionnaires. *Public Opinion Quarterly*, *58*(4), 557–568. doi:10.1086/269447

Dunkelberg, W. C., & Day, G. S. (1973). Nonresponse Bias and Callbacks in Sample Surveys. *Journal of Marketing Research*, *10*(2), 160–168. doi:10.2307/3149821

Durrant, G. B., & Steele, F. (2009). Multilevel modelling of refusal and non-contact in household surveys: evidence from six UK Government surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *172*(2), 361–381. doi:10.1111/j.1467-985X.2008.00565.x

Elliott, M. R., Little, R. J. A., & Lewitzky, S. (2000). Subsampling Callbacks to Improve Survey Efficiency. *Journal of the American Statistical Association*, *95*(451), 730–738. doi:10.1080/01621459.2000.10474261

Eurostat. (2012). *Households having access to the Internet, by type of connection.* Retrieved from http://epp.eurostat.ec.europa.eu/tgm/table.do?tab=table&init=1&language=en&pcode=tin00073&plugin=1

Eurostat. (2014). *Eurostat Data Explorer*. *Households - level of internet access*. Retrieved from http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=isoc_ci_in_h&lang=en

Eva, G., Loosveldt, G., Lynn, P., Martin, P., Revilla, M., Saris, W., & Vannieuwenhuyze, J. (2010). *Assessing the cost-effectiveness of different modes for ESS data collection*. City University London.

Fowler, F. J., Gallagher, P. M., Stringfellow, V. L., Zaslavsky, A. M., Thompson, J. W., & Cleary, P. D. (2002). Using Telephone Interviews to Reduce Nonresponse Bias to Mail Surveys of Health Plan Members. *Medical Care*, *40*(3), 190–200.

French, B. F., & Finch, W. H. (2008). Multigroup Confirmatory Factor Analysis: Locating the Invariant Referent Sets. *Structural Equation Modeling: A Multidisciplinary Journal*, *15*(1), 96–113. doi:10.1080/10705510701758349

Fricker, S., Galesic, M., Tourangeau, R., & Yan, T. (2005). An Experimental Comparison of Web and Telephone Surveys. *Public Opinion Quarterly*, *69*(3), 370–392. doi:10.1093/poq/nfi027

Fricker, S., & Tourangeau, R. (2010). Examining the Relationship Between Nonresponse Propensity and Data Quality in Two National Household Surveys. *Public Opinion Quarterly*, *74*(5), 934–955. doi:10.1093/poq/nfq064

Fuller, W. (1987). *Measurement Error Models*. New York: Wiley.

Gabler, S., & Häder, S. (2009). Die Kombination von Mobilfunk- und Festnetzstichproben in Deutschland (The Combination of Mobile and Landline Samples in Germany). In M. Weichbold, J. Bacher, & C. Wolf (Eds.), *Umfrageforschung: Herausforderungen und Grenzen*. VC Verlag für Sozialwissenschaften.

Gerbing, D. W., & Anderson, J. C. (1984). On the Meaning of within-Factor Correlated Measurement Errors. *Journal of Consumer Research*, *11*(1), 572–580.

Green, D. P., & Citrin, J. (1994). Measurement Error and the Structure of Attitudes: Are Positive and Negative Judgments Opposites? *American Journal of Political Science*, *38*(1), 256–281. doi:10.2307/2111344

Greene, J., Speizer, H., & Wiitala, W. (2008). Telephone and Web: Mixed-Mode Challenge. *Health Services Research*, *43*(1p1), 230–248. doi:10.1111/j.1475-6773.2007.00747.x

Groves, R. M. (1989). *Survey Errors and Survey Costs*. New York: Wiley.

Groves, R. M. (2006). Nonresponse Rates and Nonresponse Bias in Household Surveys. *Public Opinion Quarterly*, *70*(5), 646–675. doi:10.1093/poq/nfl033

Groves, R. M., Biemer, P. P., & Lyberg, L. E. (Eds.). (1988). *Telephone survey methodology*. John Wiley and Sons.

Groves, R. M., Cialdini, R. B., & Couper, M. P. (1992). Understanding the Decision to Participate in a Survey. *Public Opinion Quarterly*, *56*(4), 475–495. doi:10.1086/269338

Groves, R. M., & Couper, M. (1998). *Nonresponse in household interview surveys*. New Jersey: Wiley.

Groves, R. M., & Couper, M. P. (1996). Contact-Level Influences on Cooperation in Face-to-Face Surveys. *Journal of Official Statistics*, *12*(1), 63–83.

Groves, R. M., Couper, M. P., Presser, S., Singer, E., Tourangeau, R., Acosta, G. P., & Nelson, L. (2006). Experiments in Producing Nonresponse Bias. *Public Opinion Quarterly*, *70*(5), 720 –736. doi:10.1093/poq/nfl036

Groves, R. M., Fowler, F. J., Couper, M., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2010). *Survey Methodology* (2nd ed.). New Jersey: Wiley.

Groves, R. M., & Kahn, R. L. (1979). *Surveys By Telephone. A National Comparison with Personal Interviews*. New York: Academic Press.

Groves, R. M., & Lyberg, L. (2010). Total Survey Error: Past, Present, and Future. *Public Opinion Quarterly*, *74*(5), 849–879. doi:10.1093/poq/nfq065

Groves, R. M., & Peytcheva, E. (2008). The Impact of Nonresponse Rates on Nonresponse Bias. *Public Opinion Quarterly*, *72*(2), 167 –189. doi:10.1093/poq/nfn011

Groves, R. M., Presser, S., & Dipko, S. (2004). The Role of Topic Interest in Survey Participation Decisions. *Public Opinion Quarterly*, *68*(1), 2–31. doi:10.1093/poq/nfh002

Groves, R. M., Singer, E., & Corning, A. (2000). Leverage-Saliency Theory of Survey Participation: Description and an Illustration. *Public Opinion Quarterly*, *64*(3), 299–308. doi:10.1086/317990

Guo, S., & Fraser, M. W. (2010). *Propensity Score Analysis*. Thousand Oaks: Sage.

Haan, M., Ongena, Y. P., & Aarts, K. (2014). Reaching Hard-to-Survey Populations: Mode Choice and Mode Preference. *Journal of Official Statistics*, *30*(2), 355–379. doi:10.2478/jos-2014-0021

Hansen, M. H., & Hurwitz, W. N. (1946). The Problem of Non-Response in Sample Surveys. *Journal of the American Statistical Association*, *41*(236), 517–529. doi:10.1080/01621459.1946.10501894

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. doi:10.1093/ijpor/edn054

Heerwegh, D., & Loosveldt, G. (2008). Face-to-Face versus Web Surveying in a High-Internet-Coverage Population. *Public Opinion Quarterly*, *72*(5), 836–846. doi:10.1093/poq/nfn045

Heerwegh, D., & 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.

Hochstim, J. R. (1967). A Critical Comparison of Three Strategies of Collecting Data from Households. *Journal of the American Statistical Association*, *62*(319), 976–989. doi:10.2307/2283686

Holbrook, A. L., Green, M. C., & 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. doi:10.1086/346010

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, *81*(396), 945–960. doi:10.2307/2289064

Hox, J. J., & De Leeuw, E. (1994). A comparison of nonresponse in mail, telephone, and face-to-face surveys. *Quality & Quantity*, *28*(4), 329–344. doi:10.1007/BF01097014

Hox, J. J., De Leeuw, E., & Dillman, D. A. (2008). The cornerstones of survey research. In E. De Leeuw, D. A. Dillman, & J. J. Hox (Eds.), *International Handbook of Survey Methodology* (pp. 299–316). New York: Lawrence Erlbaum.

Imbens, G. W. (2004). Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review. *Review of Economics and Statistics*, *86*(1), 4–29. doi:10.1162/003465304323023651

Jäckle, A., Roberts, C., & Lynn, P. (2010). Assessing the Effect of Data Collection Mode on Measurement. *International Statistical Review*, *78*(1), 3–20. doi:10.1111/j.1751-5823.2010.00102.x

Jöreskog, K. G. (1971a). Simultaneous factor analysis in several populations. *Psychometrika*, *36*(4), 409–426. doi:10.1007/BF02291366

Jöreskog, K. G. (1971b). Statistical analysis of sets of congeneric tests. *Psychometrika*, *36*(2), 109–133. doi:10.1007/BF02291393

Kamata, A., & Bauer, D. J. (2008). A Note on the Relation Between Factor Analytic and Item Response Theory Models. *Structural Equation Modeling: A Multidisciplinary Journal*, *15*(1), 136–153. doi:10.1080/10705510701758406

Kaminska, O., McCutcheon, A. L., & Billiet, J. B. (2010). Satisficing Among Reluctant Respondents in a Cross-National Context. *Public Opinion Quarterly*, *74*(5), 956–984. doi:10.1093/poq/nfq062

Kang, J. D. Y., & Schafer, J. L. (2007). Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statistical Science*, *22*(4), 523–539. doi:10.1214/07-STS227

Kankaraš, M., Vermunt, J. K., & Moors, G. (2011). Measurement Equivalence of Ordinal Items: A Comparison of Factor Analytic, Item Response Theory, and Latent Class Approaches. *Sociological Methods & Research*, *40*(2), 279 –310. doi:10.1177/0049124111405301

Kaplan, D. (1999). An Extension of the Propensity Score Adjustment Method for the Analysis of Group Differences in MIMIC Models. *Multivariate Behavioral Research*, *34*(4), 467–492. doi:10.1207/S15327906MBR3404_4

Kaplowitz, M. D., Hadlock, T. D., & Levine, R. (2004). A Comparison of Web and Mail Survey Response Rates. *Public Opinion Quarterly*, *68*(1), 94–101. doi:10.1093/poq/nfh006

Keeter, S., Miller, C., Kohut, A., Groves, R. M., & Presser, S. (2000). Consequences of Reducing Nonresponse in a National Telephone Survey. *Public Opinion Quarterly*, *64*(2), 125 –148. doi:10.1086/317759

Kersten, H. M. P., & Bethlehem, J. (1984). Exploring and reducing the nonresponse bias by asking the basic question. *Statistical Journal of the United Nations Economic Commission for Europe*, *2*(4), 369–380.

Kim, E. S., & Yoon, M. (2011). Testing Measurement Invariance: A Comparison of Multiple-Group Categorical CFA and IRT. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*(2), 212–228. doi:10.1080/10705511.2011.557337

Kirchner, A., & Felderer, B. (2013, May 16). *The Effect of Survey Mode on Nonresponse Bias and Measurement Error: A Validation Approach*. Paper presented at the 68th Annual Conference of the American Association for Public Opinion Research, Boston.

Klausch, T., Hox, J. J., & Schouten, B. (2013a). *Assessing the mode-dependency of sample selectivity across the survey response process* (Discussion Paper No. 2013-03). The Hague, The Netherlands: Statistics Netherlands. Retrieved from http://www.cbs.nl/NR/rdonlyres/D285D803-D201-437D-99F6-3FB7C5DA9C11/0/201303x10pub.pdf

Klausch, T., Hox, J. J., & Schouten, B. (2013b). Measurement Effects of Survey Mode on the Equivalence of Attitudinal Rating Scale Questions. *Sociological Methods & Research*, *42*(3), 227–263. doi:10.1177/0049124113500480

Klausch, T., Schouten, B., & Hox, J. J. (2014). *The Use of Within-Subject Experiments for Estimating Measurement Effects in Mixed-Mode Surveys* (Discussion Paper No. 2014-06). The Hague, The Netherlands: Statistics Netherlands. Retrieved from http://www.cbs.nl/NR/rdonlyres/181793AC-94B8-4748-9C2B-E541DCF9CFB7/0/201406x10pub.pdf

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

Körmendi, E. (1988). The Quality of Income Information in Telephone and Face to Face Surveys. In R. M. Groves, P. P. Biemer, L. Lyberg, J. T. Massey, W. Nicholls, II, & J. Waksberg (Eds.), *Telephone Survey Methodology* (pp. 341–375). New York: Wiley.

Kraan, T., Van den Brakel, J., Buelens, B., & Huys, H. (2010). *Social Desirability Bias, Response Order Effect and Selection Effects in the New Dutch Safety Monitor* (Statistics Netherlands Discussion Paper No. 10004). The Hague, The Netherlands: Statistics Netherlands. Retrieved from http://www.cbs.nl/NR/rdonlyres/7DB17DBF-3FEA-47D9-9B1C-1A1F50410F81/0/201004x10pub.pdf

Kreuter, F., Müller, G., & Trappmann, M. (2010). Nonresponse and Measurement Error in Employment Research: Making Use of Administrative Data. *Public Opinion Quarterly*, *74*(5), 880 –906. doi:10.1093/poq/nfq060

Kreuter, F., Müller, G., & Trappmann, M. (2013). A Note on Mechanisms Leading to Lower Data Quality of Late or Reluctant Respondents. *Sociological Methods & Research*. doi:10.1177/0049124113508094

Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social Desirability Bias in CATI, IVR, and Web Surveys: The Effects of Mode and Question Sensitivity. *Public Opinion Quarterly*, *72*(5), 847–865. doi:10.1093/poq/nfn063

Krosnick, J. A. (1991). Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys. *Applied Cognitive Psychology*, *5*(3), 213–236. doi:10.1002/acp.2350050305

Krosnick, J. A. (1999). Survey Research. *Annual Review of Psychology*, *50*(1), 537–567. doi:10.1146/annurev.psych.50.1.537

Kwak, N., & Radler, B. T. (2002). A comparison between mail and Web surveys. Response pattern, respondent profile, and data quality. *Journal of Official Statistics*, *18*(2), 257–74.

Li, K.-H., Meng, X.-L., Raghunathan, T. E., & Rubin, D. B. (1991). Significance levels from repeated p-values with multiply-imputed data. *Statistica Sinica*, *1*(1), 65–92.

Link, M. W., & Mokdad, A. H. (2005). Alternative Modes for Health Surveillance Surveys: An Experiment with Web, Mail, and Telephone. *Epidemiology*, *16*(5), 701–704. doi:10.1097/01.ede.0000172138.67080.7f

Link, M. W., & Mokdad, A. H. (2006). Can Web and Mail Survey Modes Improve Participation in an RDD-based National Health Surveillance? *Journal of Official Statistics*, *22*(2), 293–312.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken: Wiley.

Lord, F. M., & Norvick, M. R. (1968). *Statistical theories of mental test scores*. Reading: Addison-Welsley.

Lugtig, P., Lensvelt-Mulderts, G. J. L. M., Frefrichs, R., & Greven, A. (2011). Estimating nonresponse bias and mode effects in a mixed-mode survey. *International Journal of Market Research*, *53*(5), 669–686.

Lumley, T. S. (2010). *Complex Surveys: A Guide to Analysis Using R*. Hoboken, New Jersey: Wiley.

Lyberg, L. (2013). Survey Quality. *Survey Methodology*, *38*(2), 107–130.

Lynn, P. (2013). Alternative Sequential Mixed-Mode Designs: Effects on Attrition Rates, Attrition Bias, and Costs. *Journal of Survey Statistics and Methodology*, *1*(2), 183–205. doi:10.1093/jssam/smt015

Lynn, P., & Clarke, P. (2002). Separating refusal bias and non-contact bias: evidence from UK national surveys. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *51*(3), 319–333. doi:10.1111/1467-9884.00321

Manfreda, K. L., Bosnjak, M., Berzelak, J., Haas, I., & Vehovar, V. (2008). Web surveys versus other survey modes: A meta-analysis comparing response rates. *International Journal of Market Research*, *50*(1), 79–104.

Meade, A. W., & Lautenschlager, G. J. (2004). A Comparison of Item Response Theory and Confirmatory Factor Analytic Methodologies for Establishing Measurement Equivalence/Invariance. *Organizational Research Methods*, *7*(4), 361–388. doi:10.1177/1094428104268027

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*(2), 127–143. doi:10.1016/0883-0355(89)90002-5

Mellenbergh, G. J. (1999). Measurement Models. In H. Adèr & G. J. Mellenbergh (Eds.), *Research Methodology in the Social, Behavioral & Life Sciences* (pp. 168–187). London: Sage.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. doi:10.1007/BF02294825

Millar, M. M., & Dillman, D. A. (2011). Improving Response to Web and Mixed-Mode Surveys. *Public Opinion Quarterly*, *75*(2), 249 –269. doi:10.1093/poq/nfr003

Miller, P. V., & Groves, R. M. (1985). Matching Survey Responses to Official Records: An Exploration of Validity in Victimization Reporting. *Public Opinion Quarterly*, *49*(3), 366–380. doi:10.1086/268934

Miller, T. I., Kobayashi, M. M., Caldwell, E., Thurston, S., & Collett, B. (2002). Citizen Surveys on the Web. *Social Science Computer Review*, *20*(2), 124 –136. doi:10.1177/089443930202000203

Millsap, R. E. (2011). *Statistical Approaches to Measurement Invariance*. New York: Routledge.

Millsap, R. E., & Yun-Tein, J. (2004). Assessing Factorial Invariance in Ordered-Categorical Measures. *Multivariate Behavioral Research*, *39*(3), 479–515. doi:10.1207/S15327906MBR3903_4

Mohorko, A., De Leeuw, E., & Hox, J. J. (2013a). Coverage Bias in European Telephone Surveys: Developments of Landline and Mobile Phone Coverage across Countries and over Time. *Survey Methods: Insights from the Field*. Retrieved from http://surveyinsights.org/?p=828

Mohorko, A., De Leeuw, E., & Hox, J. J. (2013b). Internet Coverage and Coverage Bias in Europe: Developments Across Countries and Over Time. *Journal of Official Statistics*, *29*(4), 609–622. doi:10.2478/jos-2013-0042

Morgan, S. L., & Harding, D. J. (2006). Matching Estimators of Causal Effects Prospects and Pitfalls in Theory and Practice. *Sociological Methods & Research*, *35*(1), 3–60. doi:10.1177/0049124106289164

Morgan, S. L., & Winship, C. (2007). *Counterfactuals and Causal Inference*. Cambridge: Cambridge University Press.

Morren, M., Gelissen, J. P. T. M., & Vermunt, J. K. (2011). Dealing with Extreme Response Style in Cross-Cultural Research: A Restricted Latent Class Factor Analysis Approach. *Sociological Methodology*, *41*(1), 13–47. doi:10.1111/j.1467-9531.2011.01238.x

Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115–132. doi:10.1007/bf02294210

Muthén, B. O., & Asparouhov, T. (2002). *Latent Variable Analysis With Categorical Outcomes: Multiple-Group And Growth Modeling In Mplus*. Mplus Web Notes. Retrieved from http://www.statmodel.com/download/webnotes/CatMGLong.pdf

Muthén, B. O., du Toit, S. H. C., & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes.* Retrieved from http://www.gseis.ucla.edu/faculty/muthen/articles/Article_075.pdf

Muthén, B. O., & Muthén, L. (2010). *IRT in Mplus*. Mplus Web Notes. Retrieved from http://www.statmodel.com/download/MplusIRT2.pdf

Nicoletti, C., & Peracchi, F. (2005). Survey response and survey characteristics: microlevel evidence from the European Community Household Panel. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *168*(4), 763–781. doi:10.1111/j.1467-985X.2005.00369.x

Olson, K. (2006). Survey Participation, Nonresponse Bias, Measurement Error Bias, and Total Bias. *Public Opinion Quarterly*, *70*(5), 737–758. doi:10.1093/poq/nfl038

Oort, F. J. (1998). Simulation study of item bias detection with restricted factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *5*(2), 107–124. doi:10.1080/10705519809540095

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, *46*(3), 598–609. doi:10.1037/0022-3514.46.3.598

Pearl, J. (2009). *Causality. Models, Reasoning, and Inference* (2nd ed.). New York: Cambridge University Press.

Peress, M. (2010). Correcting for Survey Nonresponse Using Variable Response Propensity. *Journal of the American Statistical Association*, *105*(492), 1418–1430. doi:doi:10.1198/jasa.2010.ap09485

Peytchev, A. (2009). Survey Breakoff. *Public Opinion Quarterly*, *73*(1), 1–24. doi:10.1093/poq/nfp014

Peytchev, A. (2012). Multiple Imputation for Unit Nonresponse and Measurement Error. *Public Opinion Quarterly*, *76*(2), 214–237. doi:10.1093/poq/nfr065

Presser, S., & Stinson, L. (1998). Data Collection Mode and Social Desirability Bias in Self-Reported Religious Attendance. *American Sociological Review*, *63*(1), 137–145. doi:10.2307/2657486

Raghunathan, T. E., Lepkowski, J., van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, *27*(1), 85–95.

Rosenbaum, P. R. (1987). Model-Based Direct Adjustment. *Journal of the American Statistical Association*, *82*(398), 387–394. doi:10.2307/2289440

Rosenbaum, P. R. (2002). *Observational Studies* (2nd ed.). New York: Springer.

Rosenbaum, P. R. (2010). *Design of Observation Studies*. Berlin: Springer.

Rosenbaum, P. R., & Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, *70*(1), 41–55. doi:10.1093/biomet/70.1.41

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. doi:10.1037/h0037350

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, *63*(3), 581–592. doi:10.2307/2335739

Rubin, D. B. (1977). Assignment to Treatment Group on the Basis of a Covariate. *Journal of Educational and Behavioral Statistics*, *2*(1), 1–26. doi:10.3102/10769986002001001

Rubin, D. B. (1978). Bayesian Inference for Causal Effects: The Role of Randomization. *The Annals of Statistics*, *6*(1), 34–58. doi:10.2307/2958688

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.

Rubin, D. B. (2003). Nested multiple imputation of NMES via partially incompatible MCMC. *Statistica Neerlandica*, *57*(1), 3–18. doi:10.1111/1467-9574.00217

Rubin, D. B. (2005). Causal Inference Using Potential Outcomes. *Journal of the American Statistical Association*, *100*(469), 322–331. doi:10.1198/016214504000001880

Sakshaug, J. W., Yan, T., & Tourangeau, R. (2010). Nonresponse Error, Measurement Error, And Mode Of Data Collection: Tradeoffs in a Multi-mode Survey of Sensitive and Non-sensitive Items. *Public Opinion Quarterly*, *74*(5), 907–933. doi:10.1093/poq/nfq057

Saris, W. E., & Andrews, F. M. (1991). Evaluation of Measurement Instruments Using a Structural Equation Modeling Approach. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement Errors in Surveys* (pp. 487–517). Hoboken, New Jersey: Wiley.

Saris, W. E., & Gallhofer, I. (2007). Estimation of the effects of measurement characteristics on the quality of survey questions. *Survey Research Methods*, *1*(1), 29–43.

Särndal, C.-E., & Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. Chichester: Wiley.

Sax, L. J., Gilmartin, S. K., & Bryant, A. N. (2003). Assessing Response Rates and Nonresponse Bias in Web and Paper Surveys. *Research in Higher Education*, *44*(4), 409–432. doi:10.1023/A:1024232915870

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Boca Raton: Chapman & Hall/CRC.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177. doi:10.1037/1082-989X.7.2.147

Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods*, *13*(4), 279–313. doi:10.1037/a0014268

Scherpenzeel, A. C. (2009). *Start of the LISS panel: Sample and recruitment of a probability-based Internet panel*. Tilburg: CentERdata. Retrieved from http://www.lissdata.nl/assets/uploaded/Sample%20and%20Recruitment_1.pdf

Scherpenzeel, A. C., & Saris, W. E. (1997). The Validity and Reliability of Survey Questions A Meta-Analysis of MTMM Studies. *Sociological Methods & Research*, *25*(3), 341–383. doi:10.1177/0049124197025003004

Schonlau, M., van Soest, A., Kapteyn, A., & Couper, M. (2009). Selection Bias in Web Surveys and the Use of Propensity Scores. *Sociological Methods & Research*, *37*(3), 291 –318. doi:10.1177/0049124108327128

Schonlau, M., Zapert, K., Simon, L. P., Sanstad, K. H., Marcus, S. M., Adams, J., … Berry, S. H. (2004). A Comparison Between Responses From a Propensity-Weighted Web Survey and an Identical RDD Survey. *Social Science Computer Review*, *22*(1), 128 –138. doi:10.1177/0894439303256551

Schouten, B., Cobben, F., & Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, *35*(1), 101–113.

Schouten, B., Shlomo, N., & Skinner, C. (2011). Indicators for Monitoring and Improving Representativeness of Response. *Journal of Official Statistics*, *27*(2), 231–253.

Schouten, B., van den Brakel, J., Buelens, B., van der Laan, J., & Klausch, T. (2013). Disentangling mode-specific selection and measurement bias in social surveys. *Social Science Research*, *42*(6), 1555–1570. doi:10.1016/j.ssresearch.2013.07.005

Shih, T.-H., & Fan, X. (2008). Comparing Response Rates from Web and Mail Surveys: A Meta-Analysis. *Field Methods*, *20*(3), 249–271. doi:10.1177/1525822X08317085

Shlomo, N., Skinner, C., & Schouten, B. (2012). Estimation of an indicator of the representativeness of survey response. *Journal of Statistical Planning and Inference*, *142*(1), 201–211. doi:10.1016/j.jspi.2011.07.008

Siemiatycki, J. (1979). A comparison of mail, telephone, and home interview strategies for household health surveys. *Am J Public Health*, *69*(3), 238–245. doi: 10.2105/AJPH.69.3.238

Skalland, B. (2011). An Alternative to the Response Rate for Measuring a Survey's Realization of the Target Population. *Public Opinion Quarterly*, *75*(1), 89–98. doi:10.1093/poq/nfq072

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Boca Raton, FL: Chapman and Hall.

Smyth, J. D., Dillman, D. A., Christian, L. M., & O'Neill, A. C. (2010). Using the Internet to Survey Small Towns and Communities: Limitations and Possibilities in the Early 21st Century. *American Behavioral Scientist*, *53*(9), 1423–1448. doi:10.1177/0002764210361695

Snijkers, G., Hox, J. J., & De Leeuw, E. (1999). Interviewers' Tactics for Fighting Survey Nonresponse. *Journal of Official Statistics*, *15*(2), 185–198.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology*, *91*(6), 1292–1306. doi:10.1037/0021-9010.91.6.1292

Stuart, E. A. (2010). Matching Methods for Causal Inference: A Review and a Look Forward. *Statistical Science*, *25*(1), 1–21. doi:10.1214/09-STS313

Suzer Gurtekin, Z. T. (2013). *Investigating the Bias Properties of Alternative Statistical Inference Methods in Mixed-Mode Surveys* (PhD thesis). University of Michigan, Michigan.

Thornberry, O. T., & Massey, J. T. (1988). Trends in United States Telephone Coverage Across Time and Subgroups. In R. M. Groves, P. P. Biemer, & L. E. Lyberg (Eds.), *Telephone Survey Methodology* (pp. 25–50). New York: Wiley & Sons.

Toepoel, V., & Lugtig, P. (2014). What Happens if You Offer a Mobile Option to Your Web Panel? Evidence From a Probability-Based Panel of Internet Users. *Social Science Computer Review*. doi:10.1177/0894439313510482

Tourangeau, R., Couper, M. P., & Conrad, F. (2004). Spacing, Position, and Order Interpretive Heuristics for Visual Features of Survey Questions. *Public Opinion Quarterly*, *68*(3), 368–393. doi:10.1093/poq/nfh035

Tourangeau, R., Couper, M. P., & Steiger, D. M. (2003). Humanizing self-administered surveys: experiments on social presence in web and IVR surveys. *Computers in Human Behavior*, *19*(1), 1–24. doi:10.1016/S0747-5632(02)00032-8

Tourangeau, R., Groves, R. M., & Redline, C. D. (2010). Sensitive Topics and Reluctant Respondents: Demonstrating a Link between Nonresponse Bias and Measurement Error. *Public Opinion Quarterly*, *74*(3), 413–432. doi:10.1093/poq/nfq004

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge: Cambridge University Press.

Tourangeau, R., & Smith, T. W. (1996). Asking Sensitive Questions. The Impact of Data Collection Mode, Question Format, and Question Context. *Public Opinion Quarterly*, *60*(2), 275–304. doi:10.1086/297751

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, *133*(5), 859–883. doi:10.1037/0033-2909.133.5.859

Trewin, D., & Lee, G. (1988). International Comparisons of Telephone Coverage. In R. M. Groves, P. P. Biemer, & L. E. Lyberg (Eds.), *Telephone Survey Methodology* (pp. 9–24). New York: Wiley & Sons.

Tuten, T. L., Urban, D. J., & Bosnjak, M. (2002). Internet Surveys and Data Quality: A review. In B. Batinic, U.-D. Reips, & M. Bosnjak (Eds.), *Online Social Sciences* (pp. 7–26). Seattle: Hogrefe & Huber.

US Census Bureau. (2010). *Design and methodology: American Community Survey*. Washington DC: US Census Bureau. Retrieved from http://www.census.gov/acs/www/Downloads/survey_methodology/acs_design_methodology.pdf

Van Buuren, S. (2012). *Flexible Imputation of Missing Data*. Boca Raton: CRC Press.

Van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, *76*(12), 1049–1064. doi:10.1080/10629360600810434

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(3), 1–67.

Vandenberg, R. J., & Lance, C. E. (2000). A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. *Organizational Research Methods*, *3*(1), 4 –70. doi:10.1177/109442810031002

Vannieuwenhuyze, J. (2014). On the Relative Advantage of Mixed-Mode versus Single-Mode Surveys. *Survey Research Methods*, *8*(1), 31–42.

Vannieuwenhuyze, J., & Loosveldt, G. (2013). Evaluating Relative Mode Effects in Mixed-Mode Surveys: Three Methods to Disentangle Selection and Measurement Effects. *Sociological Methods & Research*, *42*(1), 82–104. doi:10.1177/0049124112464868

Vannieuwenhuyze, J., Loosveldt, G., & Molenberghs, G. (2010). A Method for Evaluating Mode Effects in Mixed-mode Surveys. *Public Opinion Quarterly*, *74*(5), 1027 –1045. doi:10.1093/poq/nfq059

Vannieuwenhuyze, J., 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.

Voogt, R. J. J. (2004). *"I'm Not Interested". Nonresponse bias, response bias and stimulus effects in election research.* (PhD-Thesis). University of Amsterdam, Amsterdam.

Voogt, R. J. J., & Saris, W. E. (2005). Mixed Mode Designs: Finding the Balance Between Nonresponse Bias and Mode Effects. *Journal of Official Statistics*, *21*(3), 367–387.

Wagner, J. (2012). A Comparison of Alternative Indicators for the Risk of Nonresponse Bias. *Public Opinion Quarterly*, *76*(3), 555–575. doi:10.1093/poq/nfs032

Welkenhuysen-Gybels, J., Billiet, J. B., & Cambré, B. (2003). Adjustment for Acquiescence in the Assessment of the Construct Equivalence of Likert-Type Score Items. *Journal of Cross-Cultural Psychology*, *34*(6), 702–722. doi:10.1177/0022022103257070

Winship, C., & Morgan, S. L. (1999). The Estimation of Causal Effects from Observational Data. *Annual Review of Sociology*, *25*, 659–706.

Yoon, M., & Millsap, R. E. (2007). Detecting Violations of Factorial Invariance Using Data-Based Specification Searches: A Monte Carlo Study. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 435–463. doi:10.1080/10705510701301677

Appendices

Appendix A (Chapter 2)

In section 5.2, we refer to variable-specific analyses of the impact of mixed-mode follow up on single-mode selection error, which are explained in the following. The analyses on systematic level (R-indicators and distribution of *V* statistics) indicated that the difference in selection error between F2F and the three single-mode designs had predictive validity for the actual change in selection error of the mixed-mode designs. To assess this hypothesis, we estimated the simple difference in *V* indices for each benchmark variable between F2F and each of the three single-mode designs and compared these scores against the actual change induced by the mixed-mode designs (i.e., the difference in *V* between the mixed-mode and single-mode designs).

Figure A1 shows the scatterplots of this information for the three single-mode surveys. Black triangles represent the eight socio-demographic variables from table 2, whereas grey circles represent the 22 CVS variables from table 3. In nearly all cases, we found a strong positive relationship between the difference in absolute selection error of the single-mode designs and the change caused by the mixed-mode follow-up. Correlation coefficients are provided in table A1.

Table A1: Pearson correlation coefficients for the relationship of expected change[a] in absolute selection error and actual change[b] induced by the mixed-mode designs

	Telephone	Mail	Web
Socio-Demographics	.806**	.693*	.924***
CVS Variables	.025	.598**	.874***

* p<.05; ** p<.01; *** p<.001
a. Difference in *V* between F2F and the single-mode design
b. Difference in *V* between the single-mode and the mixed-mode design

These results suggest that if telephone, mail, or web exhibited stronger selection error than F2F, the mixed-mode extensions greatly contributed to reducing this error. Reversely, however, the mixed-mode extensions could also enlarge error, if F2F showed stronger error than the single modes. In figure A1, the left (negative) range from the vertical line (at zero) indicates the part of the variables showing less absolute selection error than F2F, whereas the right (positive) range indicates the part where F2F showed more selection error. The strongest relationships were found for the web mode (*r=.924* for socio-demographics and *r=.874* for CVS variables). Both negative and positive changes by the mixed-mode extension are present for socio-demographics. Recalling from table 4 and figure 2 that the *R* indicator and central tendency of the distribution of *V* were mainly unaffected by the follow-up, this analysis now provided further insights. Since selection error was both introduced and mitigated by the mixed-mode designs, selection error apparently did not change 'on average'. For the CVS variables however, it can be seen that selection error was enlarged by the mixed-mode extension, which is also reflected by *R* and figure 2. Remarkably, web showed the strongest impact of the mixed-mode extension on selection error. One potential explanation why the F2F follow-up greatly aligned selection error in the direction of F2F perhaps was the strong RR increase. Approximately half of mixed-mode web respondents were represented by F2F respondents from the follow-up (table 1, section 3.3). For mail the relationships are weaker, though still strong. Again the findings on systematic level are reflected by the variable-specific changes, where, for socio-demographics reduction in error outweighed increase and for CVS variables selection error was rather increased. The same finding applied to the socio-demographics of the telephone mode. However, for target variables this change was not found. The

difference in absolute selection error between F2F and telephone perhaps was not large enough to induce any substantial changes by the mixed-mode design.

In conclusion, this analysis shows that the F2F follow-up could both increase and decrease selection error. The direction of change could be predicted very well by the difference in selection error of the single-mode designs. By this process, absolute selection error of telephone, mail, and web became more similar to a single-mode F2F survey after the mixed-mode extension.

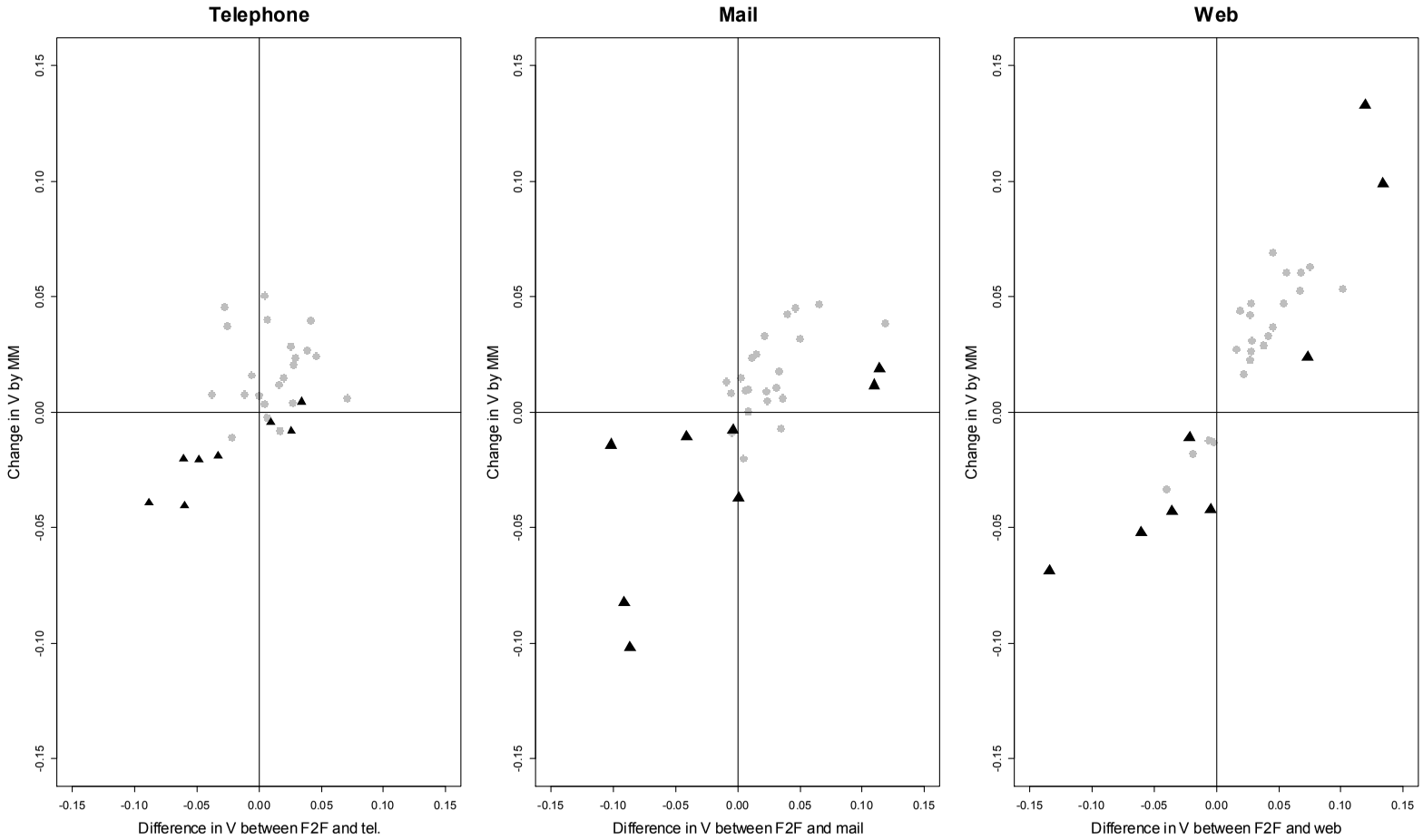


Figure A1: Scatterplots of the expected change in selection error (difference in V between F2F and a single mode design) and the actual change (difference between the single-mode and mixed-mode V). Triangles represent socio-demographics, circles represent CVS target variables.

Appendix B (Chapter 3)

Appendix B-1: Graphical illustration of equivalence testing

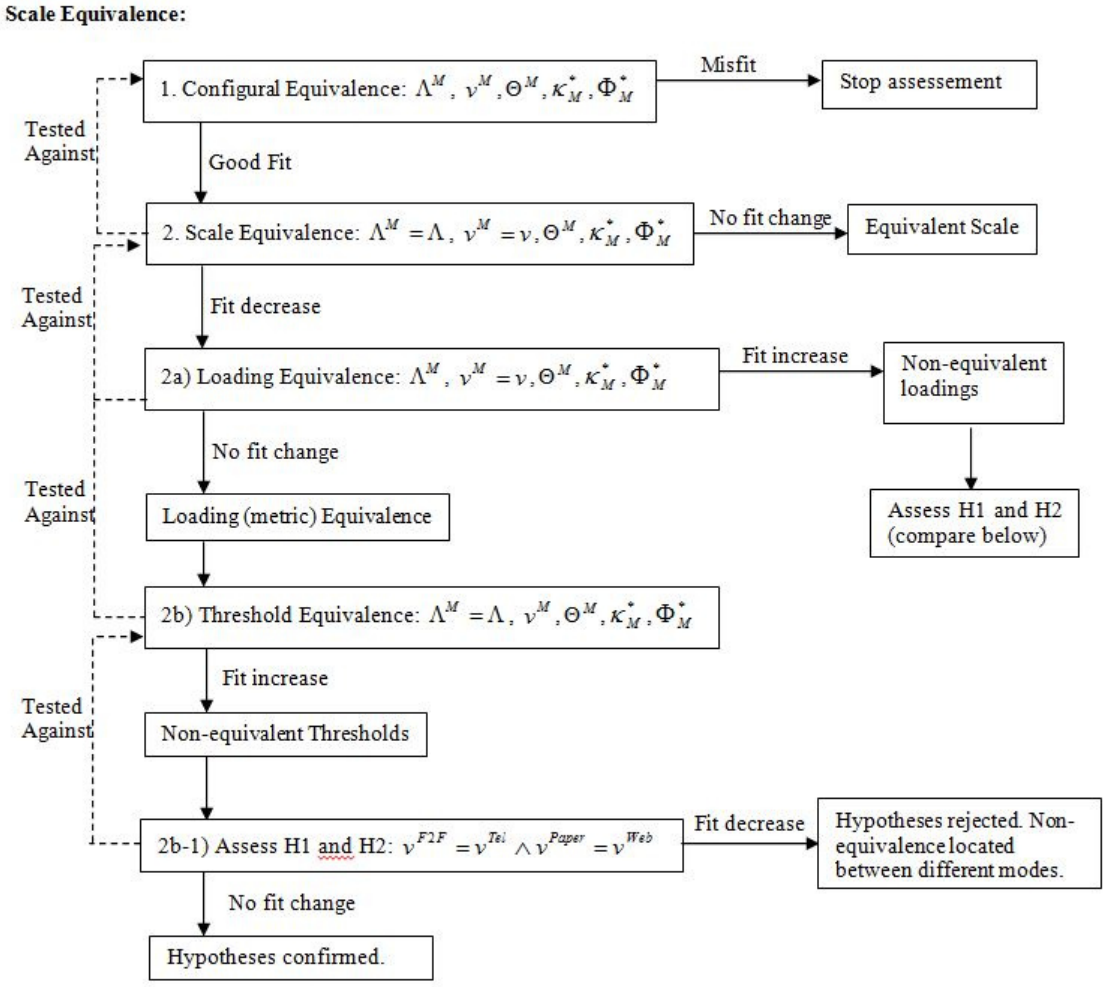
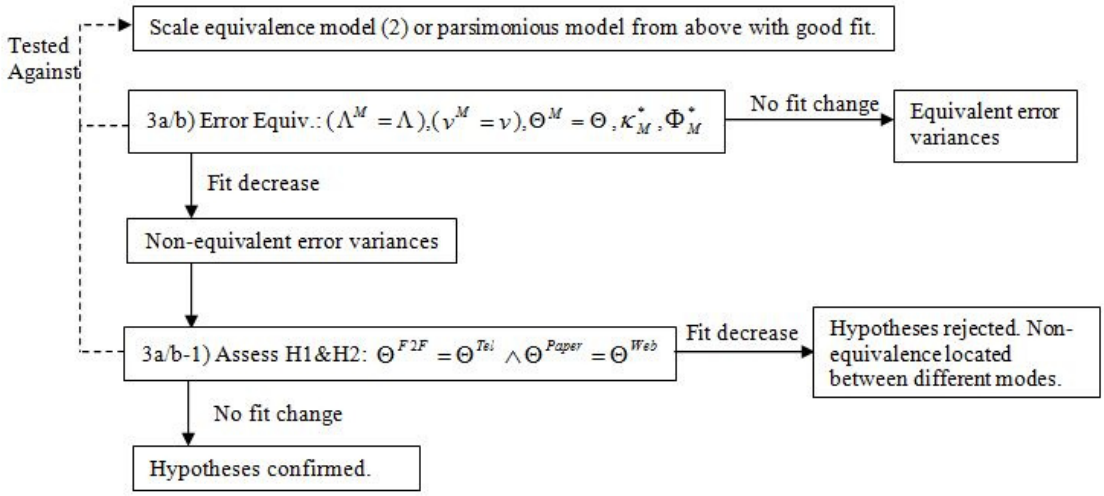
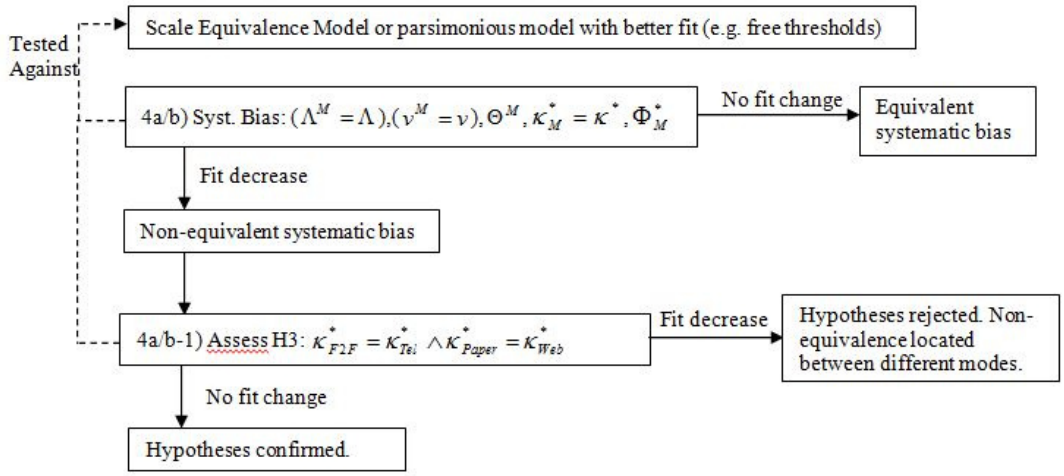


Figure B1: Graphical illustration of Testing Procedure (continued on next page)

Random Error Equivalence:



Systematic Bias Equivalence:



Systematic Variance Equivalence:

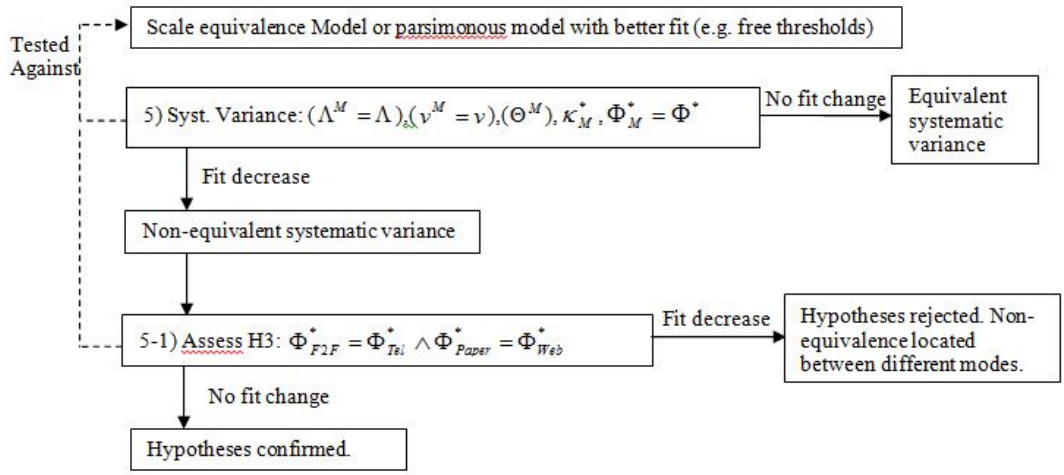


Figure B1: Graphical illustration of Testing Procedure (continued from previous page)

Appendix B-2: Results on the alternative parameterization of systematic errors

We estimated the alternative parameterization of systematic errors according to formula (3.16) in Mplus for all scales (Table A2). *T* is assumed a factor with zero means and unit variance in all conditions. *S* is a factor with free means and variances but unit constrained loadings. Testing systematic bias equivalence implied constraining means of *S* equal across mode conditions and assessing change in model fit. Testing systematic variance implied constraining variances of *S* equal. The equivalence tests matched the findings of the compound parameterization. Changes in RMSEA are very comparable to the compound parameterization. Model fit in terms of RMSEA was worse for the PV and DTO scale than the compound parameterization and slightly better for the NTP scale. Directions of means matched the compound parameterization. Additionally, systematic variance in the web condition was higher in the DTO scale, which matched the compound parameterization.

Table B1: Equivalence tests of systematic error estimated as direct effect on indicators

Model	Equivalence Type	Tested against	RMSEA	Diff. in RMSEA	Model Chi-Square	Adj. Chi² diff. test	Model df (diff.)
NTP Scale:							
2	Scale equivalence	-	.067	-	144.3	-	26
4	Systematic Bias	2	.092	+.015	277.2	106.5***	29 (3)
	Web=Paper≠F2F =Tel.	2	.064	-.003	142.7	8.6 (n.s.)[b]	28 (2)
5	Systematic Variance	2	.070	+.003	169.8	23.3 (n.s.)[b]	29 (3)
PV Scale:							
2	Scale equivalence	-	.063	-	239.2***	-	50
4	Systematic Bias	2	.072	+.009	315.9***	52.9*[b]	53 (3)
	Web=Paper≠F2F =Tel.	2	.052	-.009	185.8***	4.2 (n.s.)	52 (2)
5	Systematic Variance	2	.061	-.002	176.0***	4.1 (n.s.)	53 (3)
DTO Scale:							
2	Scale equivalence	-	.034	-	64.4***	-	30
4	Systematic Bias	2	.040	+.060	88.0***	17.2*[b]	33 (3)
	Web=Paper≠F2F =Tel.	2	.026	-.008	53.8**	3.6 (n.s.)	32 (2)
5	Systematic Variance	2	.048	+.014	108.2***	25.2*[b]	33 (3)
	Web=Paper≠F2F =Tel.	2	.049	+.015	109.6***	22.3*[b]	32 (2)
	Paper=Cati=Capi≠Web	2	.032	-.002	65.2***	5.5 (n.s.)	32 (2)

For sample sizes N see tables in main text.
* p<.05, ** p<.01, *** p<.001
'≠' denotes free parameters for two modes, '=' denotes fixed parameters for two modes.
a. Meaningful change of RMSEA criterion (i.e. >.01)
b. Oort adjustment of critical value resulted in a lower significant level. In all other cases adjustment did not change level of significance
c. Effect / Significance did not hold to cross-validation in both split half samples.

Appendix C (Chapter 4)

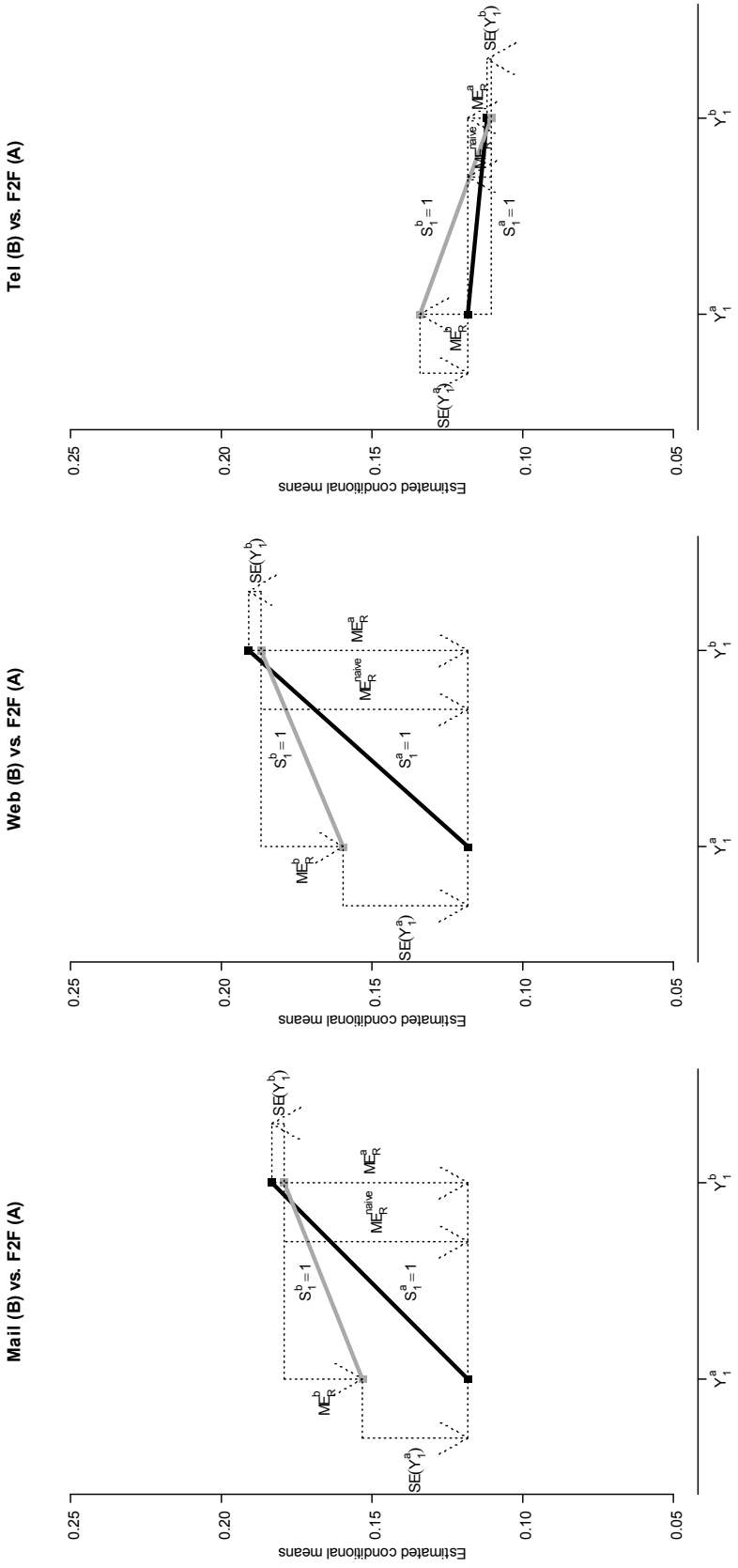


Figure C1: Interaction diagrams for the proportion of individuals feeling unsafe show significant differences in conditional *MEs* for mail (p<.05) and web (p=.066) (Interaction effect). All (adjusted) estimates based on the WSCGD (F2F respondents: black; mail, web, and telephone respondents: grey).

Appendix D (Chapter 5)

Table D1: Overview on items with type of target statistics and item nonresponse

	No. cat.	Target Statistic	% item nonresponse (DK / refusal) F2F (n=1,048)	Telephone (n=735)	Mail (n=857)	Web (n=497)
'Social quality neighbourhood'[a]*:*						
State of roads, walkways, squares	5	% (completely) agree	0.0	0.0	4.3	0.6
Good playgrounds for children	5	"	4.8	3.1	8.8	4.6
Good provisions for the young	5	"	9.3	7.2	15.2	9.9
People know each other well	5	"	0.6	0.4	5.6	1.4
People treat each other well	5	"	0.5	0.1	5.1	0.6
Nice neighbourhood with solidarity	5	"	1.0	0.1	5.8	0.6
Feel at home with people	5	"	0.0	0.1	4.7	1.0
Have a lot of contact with people	5	"	0.0	0.0	4.2	0.4
Satisfied with population compos.	5	"	0.3	0.1	4.9	0.8
'Neighbourhood problems'[b]*:*						
Plastering on walls and/or buildings	3	% frequently / sometimes	0.3	0.4	8.5	5.2
Harassment by groups of young	3	"	0.4	0.3	4.6	2.4
Drunken people on the streets	3	"	0.7	0.7	7.4	6.2
Unpleasant people on the streets	3	"	1.5	1.5	15.2	12.5
Junk on the streets	3	"	0.3	0.1	2.9	0.4
Dog excrements on the streets	3	"	0.1	0.3	3.2	1.0
Destruction of telephone cells, etc.	3	"	4.5	4.2	17.9	14.9
Drug problems	3	"	2.0	1.4	21.2	16.9
Summary ratings neighbourhood:						
Insecurity feeling in general[c]	2	% yes	0.0	0.1	4.6	2.8
Frequency insecurity feeling[b]	3	% frequently / sometimes	0.0	0.0	2.9	0.0
Insecurity feeling in neighbourhood[c]	2	% yes	0.0	0.1	4.1	2.2
Frequency insecurity feeling[b]	3	% frequently / sometimes	0.0	0.0	0.7	0.0
Quality of Life Rating[d]	10	Mean (1 to 10)	0.3	0.5	3.9	1.8
Police evaluation						
Police contact (past 12 mth.)[c]	2	% yes	0.1	0.0	5.3	2.2
Evaluation Police Contact[e]	5	% (very) satisfied	0.0	0.5	1.9	2.1
Evaluation Police Performance[e]	5	"	7.6	3.9	20.5	11.9
Victimization						
No. of victimizations[f] (past 12 mth.)	Count	Mean	0.0	0.0	0.0	0.0
Victim of crime[f] (past 12 mth.)	2	% yes	0.0	0.0	0.0	0.0
Summary Indices (scales)						
Social cohesion[f]	Scale	Mean (0 to 10)	1.7	0.8	7.9	2.2
Neighbourhood Nuisances	Scale	Mean (0 to 10)	4.0	2.9	27.3	23.7
Neighbourhood Deterioration	Scale	Mean (0 to 10)	4.9	4.6	21.9	17.7

a. Five rating scale answer categories from 'Completely disagree' to 'Completely Agree'.
b. Three rating scale answer categories: 'Happens Frequently', 'Sometimes', and 'Rarely or never'.
c. Dichotomous (yes / no)
d. Question 'Do you sometimes feel insecure?' with two answer categories: yes, no
e. Rating of satisfaction from 1 (very satisfied) to 5 (very unsatisfied)
f. Victimization variables are aggregated based on multiple questions about victimization in the past year (count / dichotomous)
g. Ten point rating scale from 1 (very low) to 10 (very high).

English Summary

Mixed-mode data collection designs have strongly gained in prominence in world-wide survey research in the course of the past decade. "Mixed-mode surveys" combine more than one mode of administration in the same survey project. Prominent examples include so-called sequential and concurrent mixed-mode designs. Sequential mixed-mode designs offer nonrespondents or non-covered individuals in one mode (e.g., web self-administered questionnaires) the option to reply in at least one other mode (e.g., face-to-face interviewing). Concurrent designs provide respondents with a choice among multiple modes of administration at the same point in time (e.g., a choice between a web and a paper self-administered questionnaire).

Mixed-mode data collection has two primary objectives. On the one hand, survey practitioners apply mixed-mode designs to increase response and coverage rates of single-mode data collection designs, which often cover the population to lesser extents and feature lower response rates than multiple modes administered jointly. In doing so, it is assumed that mixed-mode surveys may lead to smaller selection error of survey estimates. On the other hand, mixed-mode data collection saves on survey costs compared to traditional single-mode surveys, such as face-to-face or telephone. This goal is reached, for example, by administering the inexpensive web mode as a single mode of interviewing first and offer more costly modes, such as telephone or face-to-face, only to the share of nonrespondents in the web survey.

A potential problem of mixed-mode surveys is that modes may measure questions with differing extent of systematic and random measurement error. It is therefore desirable that answers provided under different modes are 'measurement equivalent'

suggesting that measurement error is equivalent across modes and that the same respondent in expectation provides equivalent answers under different modes. A problem in mixed-mode research is that mode differences in selection error (selection effects) are confounded with mode differences in measurement error (measurement effects). Any analysis of selection and measurement error differences between modes has to deal with this confounding problem.

This dissertation addresses several methodological issues relevant for an informed design of mixed-mode data collection surveys. The reader receives an overview on the available and new methods for evaluating mode effects on selection and measurement error in single-mode and mixed-mode designs. Each chapter addresses the evaluation of one of these error components (chapters 2 to 4) or both (chapter 5) on the basis of experimental data using different statistical approaches.

Chapter 2 compares the extent of selection error evoked by the four major contemporary modes of data collection (face-to-face [F2F], telephone, mail, and web) and three sequential mixed-mode designs (telephone, mail, and web with F2F follow-up) for the case of the Dutch Crime Victimization Survey (CVS). Socio-demographic characteristics and target variables from the CVS serve as benchmark variables. A special two-wave experimental design, which is also applied as data basis for the subsequent chapters, allows studying design differences in selection error on CVS variables independently from differences in measurement error, thus controlling for the confounding problem. This design was implemented as the so-called MEPS experiment (in Dutch: mode effecten in persoons enquêtes / mode effects in person surveys) in 2011 by Statistics Netherlands. Despite large differences in response rates, only small or no differences in selection error between the four single-mode designs are found on both types of variables. There are mixtures of cases where error is enlarged or mitigated in the mixed-mode designs despite the fact that mixed-mode yields large response increases. These results suggest that selection error is less important than often believed in single and mixed-mode surveys.

Chapter 3 addresses mode differences in measurement error (measurement equivalence). It is shown how to apply ordinal confirmatory factor analysis for multiple groups (MCFA) to assess equivalence of scale, random errors and systematic

(nonrandom) errors of attitudinal questions from the CVS and the European Social Survey (ESS) across the four major modes in the MEPS experiment. Consistent with theoretical expectations, interviewer and self-administered surveys measure all assessed questions on systematically different scales, with different systematic bias, and with differing extents of random error. These measurement effects are absent when comparing mail with web or face-to-face with telephone. The self-administered modes appeared more efficient exhibiting higher indicator reliabilities (smaller random error) than the interviewer modes. It is concluded that modes impact consistent measurement effects that affect multiple questions in similar ways suggesting that question-specific content is less important in influencing the occurrence of measurement effects.

The analysis presented in chapter 3 additionally addresses the confounding problem. By simple comparisons of MCFA models across mode-specific response sets, confounding of the observed measurement effects with selection effects would not be excludable. For example, the finding that self-administered modes exhibit smaller random error may not represent a 'true' measurement effect, if respondents, who provide more accurate answers, self-select to the mail and the web modes more frequently than in interviewer-administered modes. Such selection effects have to be controlled in analyses of measurement equivalence. In chapter 3, a method called inverse propensity score weighting is applied in the estimation of the MCFA models. The propensity score is an estimated probability for a respondent to self-select into a specific mode. Propensity score weighting to the sample is applied using eight socio-demographic characteristics from the national register of The Netherlands.

Under the assumption that units are missing at random (MAR) in the mode-specific samples, or, put differently, that the socio-demographic variables explain the distributional differences of target variables between respondents and nonrespondents in each mode, the estimation of measurement effects on population level is possible. A relevant concern is that this assumption may not be true under many practical circumstances. The availability of eight auxiliary variables applied in chapter 3, can already be described as a relatively rich data basis. Survey researchers outside of statistical offices often have far less sampling frame information available in practice. In addition, socio-demographic variables are typically relatively weakly associated with response mechanisms (regardless of modes) or target variables. In the case of the MEPS

experiment, the relative differences on socio-demographic variables between modes were small (chapter 2). There is therefore reason to assume that nonresponse adjustment on socio-demographics may be insufficient to identify all measurement effects.

The objective of **chapter 4** is to address this deficiency by making available stronger auxiliary information for the adjustment of selectivity in relative comparisons of modes. A new framework to estimate measurement effects using so-called "within-subject designs" (WSD) is presented. WSDs approach the same sample by two different modes at two subsequent points in time (also called, "cross-over designs"). WSDs allow adjusting nonresponse across repeated measurement occasions by repeatedly measured target variables implying weaker MAR assumptions than adjusting only for socio-demographics. This argument is based on the conjecture that repeatedly measured target variables often are strongly related.

The measurement effects that can be estimated are conditioned on a response group in the WSD (e.g., respondents under a face-to-face mode) and are called "conditional measurement effects" for this reason. It is discussed that the conditional measurement effects that can be estimated are useful for evaluating measurement equivalence in two common mixed-mode design situations: the switch of a single-mode survey to a different single-mode survey and the introduction of a sequential mixed-mode survey. In both situations a "measurement benchmark" mode is selected (also called preferred mode), against which measurement effects can be evaluated. A practical goal of a WSD is to estimate conditional measurement effects for different candidate mixed-mode designs and decide on the need to nullify measurement effects by appropriate mixed-mode redesign (so-called unified mode design) towards the measurement benchmark.

Chapter 4 also provides a detailed discussion on the relevant assumptions needed for estimating conditional measurement effects from WSD data, in particular (a) time-stability of target variables and response probabilities and (b) the equivalence of measurement occasions. In extensions of simple within-subject designs, an independent control group, to which the same mode is assigned at both occasions, is useful to test these assumptions and adjust for time-instability. The estimation of measurement effects is illustrated for key statistics from the CVS.

Chapters 2 to 4 considered mode effect on either selection or measurement error while keeping the other component constant. **Chapter 5**, finally, presents an integrated approach in which the trade-off of measurement and selection error is evaluated. To do so, the benchmark mode assumption introduced in chapter 4 is formalized. A benchmark can have two components, a measurement mode benchmark and a selection mode benchmark. The measurement mode benchmark is assumed to provide optimal or "preferred" measurements for the survey variable of interest, whereas the selection benchmark is assumed to evoke smallest selection error across variables (put differently, it features smallest variance in response propensities). Typically, measurement and selection benchmarks are taken from the same single-mode survey. This benchmark consequently is called the single-mode benchmark (e.g., face-to-face). When both benchmarks are taken from different modes, a mixture benchmark is created called the 'hybrid-mode' benchmark (e.g., measurements from web, selection mechanism from face-to-face).

Consequently, two types of effects are defined: single-mode and mixed-mode effects. Each effect represents a difference in (systematic) survey error between a response sample estimate and the benchmark estimate (e.g., single-mode benchmark response sample estimate for a survey variable). The single-mode total effect represents a total difference of a single-mode survey and the benchmark, whereas the mixed-mode total effect represents the difference of the mixture estimate from a (sequential) mixed-mode survey and the benchmark. Single-mode and mixed-mode total effects can be decomposed into measurement and selection effect components against the benchmark, respectively. The key idea is that the single-mode effect needs to be larger in absolute sense than the mixed-mode effect to motivate the use of a mixed-mode survey. Furthermore, the source of the total effect against the benchmark may give survey designers an indication, whether measurement effects may be preventable by questionnaire redesign (e.g., unified mode design) or if a selection effect underlies observed total effects, which is not preventable by questionnaire design.

The estimation of all effects against a single-mode face-to-face benchmark and a hybrid-mode web/face-to-face benchmark is discussed using re-interview data from WSDs introduced in chapter 4. Empirical analyses on a range of CVS variables for the three sequential mixed-mode designs in the MEPS experiment (web, mail, telephone

followed-up in face-to-face) suggest that conclusions on the optimal mixed-mode design strongly depend on the underlying measurement benchmark assumption. This conjecture implies that the primary components of total survey error against the benchmarks are measurement effects, whereas selection effects are small and statistically insignificant in all cases. In the case of a single-mode face-to-face benchmark, a face-to-face follow up to telephone, mail, or web can reduce total effects (and measurement effects). When web measurements are set as preferred mode, the face-to-face follow-up increase total and measurement effects, so that a design involving only mail and/or web but not a face-to-face follow-up is recommended.

Chapter 6 concludes the dissertation with a wider discussion on the empirical findings and relevant steps in reproducing similar studies of mode effects on survey error. In particular, the following conclusions are drawn:

1. Measurement effects between interviewer and self-administered modes are the dominant component of total mode effects in the CVS. Depending on the measurement benchmark assumption, it is either desirable to use only interviewer modes or only self-administered modes in mixed-mode designs of the CVS. An advantage of self-administration may be that many attitudinal questions showed smaller random error (higher item reliability) than in interviewer-administered modes.
2. The single-mode designs in the CVS show small differences in selection error on socio-demographic characteristics and only very small and insignificant differences in target variables. An overall reduction of selection error by a sequential mixed-mode follow up is not found in empirical data, suggesting that one of the classical motivations behind mixed-mode surveys cannot be confirmed for the CVS. Using inexpensive single-mode data collection designs such as web or mail may be considered for this reason, but has to be evaluated against the desired measurement benchmark.
3. Replicating the empirical findings is strongly desirable, because the dependence of survey errors on topics, questionnaires, and populations may have strong impact on the results.

4. Re-interview designs, such as the within-subject design, may offer very valuable auxiliary data for decomposing the confounded measurement and selection effects in mixed-mode surveys.
5. The choice of preferred measurement and selection benchmark modes is an important step in mixed-mode analyses to give mode effects on systematic survey error a substantial interpretation.
6. An advantage of analyses of random and systematic error variance is that a measurement benchmark mode is not needed, but such analyses do not consider bias and therefore should not be relied upon alone.
7. Additional analyses of selection error can try to extrapolate benchmark estimates to the population and use indicators of survey representativeness to support choice of selection benchmarks.
8. Experimental split-ballot mode comparisons are important steps in designing mixed-mode surveys, but the conditional single-mode measurement and selection effects observed in these designs may not be equal to the effects observed in sequential or concurrent mixed-mode surveys. Therefore, evaluation of mode effects ultimately has to consider mixed-mode data.

In concluding, this dissertation calls for a stronger methodological approach to the field of mixed-mode research. It is argued that future research should primarily focus on two aspects. First, the statistical theory of mixed-mode survey error should be developed further. Second, it is underlined that a next important step is developing methodology for the adjustment of measurement effects in mixed-mode data. The re-interview approach developed in this dissertation may offer valuable auxiliary data to be used in this endeavour.

Nederlandse Samenvatting (Summary in Dutch)

In het afgelopen decennium zijn onderzoeksontwerpen (designs) die gebruik maken van **mixed-mode dataverzameling** wereldwijd sterk in populariteit toegenomen. "Mixed-mode surveys" combineren meer dan één manier van dataverzameling in dezelfde survey. Toonaangevende vormen zijn sequentiële en gelijktijdige ("concurrent") mixed-mode designs. Sequentiële mixed-mode designs bieden non-respondenten of personen zonder toegang tot een specifieke mode (bijvoorbeeld survey via het internet) de optie om te antwoorden in ten minste één andere mode (bijvoorbeeld een aan-huis interview). Gelijktijdige (concurrent) designs bieden respondenten de keuze tussen verschillende modes op hetzelfde tijdstip (bijvoorbeeld de keuze om een vragenlijst via het internet of via een postale papieren versie in te vullen).

Mixed-mode dataverzameling heeft twee primaire doelen. Aan de ene kant willen onderzoekers mixed-mode designs toepassen om het responspercentage en het dekkingspercentage van single-mode designs te verhogen. In single-mode designs wordt de populatie vaak in mindere mate gedekt en ook liggen de responspercentages lager dan wanneer mixed-mode designs worden toegepast. Hierbij wordt aangenomen dat wanneer mixed-mode designs worden toegepast, dit zal leiden tot een kleinere selectiefout van de steekproefschattingen. Aan de andere kant bespaart mixed-mode waarneming geld in vergelijking met de meer traditionele single-mode survey zoals aan-huis interviews of telefonische interviews. Deze besparing wordt bereikt door bijvoorbeeld als eerste de goedkopere internet-mode als single-mode aan te bieden om daarna de duurdere modes zoals telefonische of aan-huis interviews aan te bieden aan het resterende aandeel niet-respondenten in de internet-mode.

Een potentieel probleem met mixed-mode surveys is dat de verschillende modes tot een verschillende mate van systematische en toevallige meetfouten kunnen leiden. Het is daarom wenselijk dat antwoorden die in de verschillende modes zijn gegeven 'meet-equivalent' zijn, wat betekent dat meetfouten equivalent zijn tussen de verschillende modes en dat gemiddelde genomen dezelfde respondent dezelfde antwoorden geeft in verschillende modes. Een probleem met mixed-mode onderzoek is dat de verschillen tussen modes in selectiefouten (selectie-effecten) en de verschillen tussen modes in meetfouten (meet-effecten) moeilijk van elkaar te scheiden zijn. Deze effecten zijn met elkaar vermengd ('confounded'). Iedere analyse naar verschillen in selectie- en meetfouten tussen modes moet een methode vinden om deze vermenging aan te pakken.

Deze dissertatie behandelt verschillende methodologische problemen die relevant zijn voor een 'geïnformeerd' ontwerp van mixed-mode surveys. De lezer krijgt een overzicht van de beschikbare en nieuwe methoden om de effecten van dataverzamelings modes op de selectie- en meetfouten in single-mode en mixed-mode designs te evalueren. Ieder hoofdstuk bespreekt de evaluatie van één van deze fout componenten (hoofdstuk 2 t/m 4) of van beide (hoofdstuk 5) op basis van experimentele data, waarbij gebruik wordt gemaakt van verschillende statistische benaderingen.

Hoofdstuk 2 vergelijkt de omvang van selectiefouten veroorzaakt door vier belangrijke, hedendaagse dataverzamelingsmodes (interview aan-huis, telefonisch, postaal schriftelijke vragenlijst, internet vragenlijst) en drie sequentiële mixed-mode designs (telefoon, schriftelijk en internet, ieder met aan-huis follow-up) voor de Nederlandse Veiligheidsmonitor (VM). Dit is de nationale "Crime Victimization Survey" (CVS) voor Nederland. Sociaal-demografische kenmerken en doelvariabelen van de VM worden als ijkpunt gebruikt. Een speciaal experimenteel design met twee golven, dat ook wordt gebruikt als basis voor de volgende hoofdstukken, zorgt ervoor dat verschillen tussen designs in selectiefouten op VM variabelen onafhankelijk van verschillen in meetfouten onderzocht kunnen worden. Deze experimentele opzet maakt het mogelijk de vermenging (confounding) tussen selectie en meetfouten te ontwarren, waardoor er onderscheid gemaakt kan worden tussen de twee. Dit design is geimplimenteerd als het MEPS experiment (mode effecten in persoons-enquêtes) dat in 2011 door het Centraal Bureau voor de Statistiek (CBS) is uitgevoerd. Ondanks grote verschillen in responspercentages werden er alleen kleine, verwaarloosbare verschillen

gevonden in selectiefouten tussen de vier single-mode designs voor beide typen variabelen. Er zijn enkele gevallen waarbij fouten ofwel groter zijn dan wel kleiner zijn in de mixed-mode designs ondanks dat deze mixed-mode designs hogere responspercentages hebben. Deze resultaten suggereren dat selectiefouten minder belangrijk zijn dan vaak wordt gedacht in single en mixed-mode surveys.

Hoofdstuk 3 bespreekt mode verschillen in meetfouten (meet-equivalentie). Er wordt getoond hoe men ordinale confirmatieve factoranalyse voor meerdere groepen (MCFA) toe kan passen om equivalentie van schaal, toevallige fouten en systematische (niet-toevallige) fouten van attitude-vragen van de VM en de European Social Survey (ESS) te onderzoeken op de vier survey modes in het MEPS experiment. Consistent met de theoretische verwachtingen, meten interviews en surveys die zelf worden ingevuld alle vragen met systematisch verschillende schalen, met verschillende systematische bias, en met verschillende mate van toevallige meetfouten. Deze meeteffecten zijn er niet wanneer postale surveys op papier worden vergeleken met vragenlijsten via internet, of wanneer persoonlijk interviews aan huis worden vergeleken met telefonische interviews. De modes waarbij respondenten zelf de vragenlijst in moeten vullen waren efficiënter door een grotere betrouwbaarheid van de indicatoren (minder toevallige fouten) vergeleken met de modes waarbij de respondent werd geïnterviewd. De conclusie is dat modes invloed hebben op consistente meeteffecten die vervolgens op vergelijkbare wijze invloed hebben op meerdere vragen. Dit suggereert dat vraag-specifieke inhoud minder belangrijk is in het optreden van meeteffecten.

Hoofdstuk 3 betreft methodologische oplossingen voor de vermenging van meet- en selectie effecten. Door eenvoudige vergelijkingen van MCFA modellen tussen de verschillende mode-specifieke datasets kan het onderscheid tussen de meet- en selectie effecten niet goed worden gemaakt. Bijvoorbeeld, de bevinding dat modes waarbij respondenten zelf de vragenlijst invullen tot minder toevallige fouten zou leiden, zou daarom niet een echt meeteffect hoeven te zijn, als respondenten die meer nauwkeurige antwoorden geven eerder voor modes kiezen zoals een vragenlijst via de post of via het internet dan voor een mode waarbij zij worden geïnterviewd. In meet equivalentie analyses moet worden gecorrigeerd voor zulke selectie-effecten. In hoofdstuk 3 wordt een methode die “inverse propensity score weging” heet toegepast in de schatting van de MCFA modellen. De propensity score is de geschatte kans voor een respondent om

zichzelf in een bepaalde mode te selecteren. Deze modellen worden alle naar de steekproef gewogen gebruikmakend van acht sociaal-demografische kenmerken uit het nationale registrater van Nederland.

Onder de aanname dat respondenten in de mode-specifieke steekproeven conditioneel op de achtergrond kenmerken willekeurig ontbreken ("missing at random", MAR), of, anders gezegd, dat de sociaal-demografische variabelen de verschillen in de verdelingen van de doelvariabelen tussen respondenten en niet-respondenten in iedere mode verklaren, is de schatting van de meeteffecten op populatieniveau mogelijk. Een zorg is dat in veel praktische omstandigheden niet aan deze aanname wordt voldaan. De beschikbaarheid van acht aanvullende variabelen toegepast in hoofdstuk 3 kan worden gezien als een relatief rijke dataset. Onderzoekers die vragenlijsten gebruiken en niet van een statistisch bureau komen hebben in de praktijk vaak veel minder informatie over het steekproefkader tot hun beschikking. Daarnaast zijn sociaal-demografische variabelen vaak zwak gerelateerd aan responsmechanismen (onafhankelijk van modes) of aan doelvariabelen. In het geval van het MEPS experiment waren de relatieve verschillen op sociaal-demografische variabelen tussen modes klein (hoofdstuk 2). Er is daarom reden om aan te nemen dat correctie voor non-respons op basis van sociaal-demografische variabelen onvoldoende kan zijn om alle meeteffecten te vinden.

Het doel van **hoofdstuk 4** is om deze tekortkoming aan te pakken door krachtige aanvullende informatie in te zetten bij de correctie voor selectiviteit in relatieve vergelijkingen van modes. Een nieuw kader om meeteffecten te schatten door gebruik te maken van "within-subject designs" (WSD) wordt gepresenteerd. WSDs benaderen dezelfde steekproef met twee verschillende modes op twee tijdstippen (ook wel "cross-over designs" genoemd). Met WSDs kan er worden gecorrigeerd voor non-respons over herhaalde metingen door gebruik te maken van meerdere malen gemeten doelvariabelen, waardoor MAR aannames zwakker zijn dan wanneer er alleen wordt gecorrigeerd voor sociaal-demografische variabelen. Dit is gebaseerd op de veronderstelling dat herhaald gemeten doelvariabelen vaak sterk gecorreleerd zijn.

De meeteffecten die kunnen worden geschat worden geconditioneerd op een responsgroep in de WSD (bijvoorbeeld respondenten die een aan-huis interview hebben gehad), en worden daarom "conditionele meeteffecten" genoemd. De conditionele

meeteffecten die kunnen worden geschat zijn bruikbaar om meetequivalentie te evalueren in twee gebruikelijke mixed-mode designs: het overschakelen van een single-mode survey naar een andere single-mode survey, en de introductie van een sequentiële mixed-mode survey. In beide situaties wordt er een vergelijkings- of ijkings-mode ("measurement benchmark") geselecteerd, ook wel voorkeurs-mode ("preferred mode") genoemd, waartegen meeteffecten kunnen worden geëvalueerd. Een praktisch doel van een WSD is om conditionele meeteffecten voor verschillende kandidaat mixed-mode designs te schatten en te beslissen of het nodig is meeteffecten tussen de voorkeursmode en een andere kandidaat mode te minimaliseren. Dit kan door een geschikt mixed-mode herontwerp worden bewerkstelligd (een zogenaamd "unified mode design").

Hoofdstuk 4 voorziet in een gedetailleerde discussie over de relevante aannames die nodig zijn om conditionele meeteffecten van WSD data te schatten, met name (a) tijd-stabiliteit van doelvariabelen en responskansen en (b) de equivalentie van meetmomenten. In uitbreidingen van eenvoudige tussen-personen designs kan een onafhankelijke controlegroep die beide modes heeft ingevuld op beide tijdstippen bruikbaar zijn om deze aannames te testen, en te corrigeren voor tijd-instabiliteit. De schatting van meeteffecten wordt geïllustreerd aan de hand van een aantal belangrijke statistieken van de VM.

Hoofdstuk 2 tot en met 4 bespreken mode-effecten op ofwel selectie- ofwel meetfouten, waarbij de andere component steeds gelijk wordt gehouden. **Hoofdstuk 5**, tenslotte, bevat een geïntegreerde aanpak, waarin de afweging tussen selectie- en meetfouten wordt geëvalueerd. Om dit te kunnen doen wordt de aanname van een voorkeursmode (benchmark), geïntroduceerd in hoofdstuk 4, geformaliseerd. Een benchmark kan twee componenten hebben, één voor de meting ("measurement mode benchmark") en één voor de selectie ("selection mode benchmark"). De measurement mode benchmark wordt verondersteld te leiden tot optimale of "geprefereerde" metingen voor de survey variabele waarin men in is geïnteresseerd, terwijl de selection benchmark wordt verondersteld te leiden tot de kleinste selectiefouten over alle variabelen (anders gezegd, het leidt tot de kleinste variantie in respons propensities). Doorgaans worden beide benchmarks bepaald aan de hand van dezelfde single-mode survey. Deze benchmarks worden daarom de single-mode benchmark genoemd (bijvoorbeeld een aan-huis of face-to-face interview). Wanneer de benchmarks uit twee

verschillende modes komen, ontstaat er een gemengd criterium, ofwel een 'hybrid-mode' benchmark (bijvoorbeeld, meeteffecten van de internet-mode, selectiemechanisme van de aan-huis mode).

Als gevolg hiervan worden twee typen effecten gespecificeerd: single-mode en mixed-mode effecten. Elk effect representeert een verschil in (systematische) surveyfouten tussen een steekproefschatting en de benchmarkschatting (bijvoorbeeld een single-mode benchmark steekproefschatting voor een survey variabele). Het totale single-mode effect representeert een totaal verschil van een single-mode survey en de benchmark, terwijl het mixed-mode totale effect een verschil representeert van de gecombineerde schatting van een (sequetiële) mixed-mode survey en de benchmark. Single-mode en mixed-mode totale effecten kunnen respectievelijk worden opgesplitst in meet- en selectie-effect componenten vergeleken met de benchmark. De kern is dat het single-mode effect groter moet zijn in absolute waarde dan het mixed-mode effect, om het gebruik van een mixed-mode survey aan te raden. Daarnaast kan de bron van het totale effect tegen het criterium survey, ontwerpers een idee geven of meeteffecten voorkomen zouden kunnen worden door de vragenlijst te herontwerpen (bijvoorbeeld een unified-mode design), of dat een selectie-effect onderdeel is van het totale geobserveerde effect, wat niet voorkomen kan worden door een herontwerp van de vragenlijst.

De schatting van alle effecten tegen een single-mode aan-huis-interview benchmark en een hybrid-mode internet/aan-huis-interview benchmark wordt besproken, waarbij gebruik wordt gemaakt van re-interview data van de WSDs die zijn geïntroduceerd in hoofdstuk 4. Empirische analyses op verschillende VM variabelen voor de drie sequentiële mixed-mode designs in het MEPS experiment (internet, post, telefoon, allen opgevolgd door aan-huis) suggereren dat conclusies over het optimale mixed-mode design sterk afhankelijk zijn van de onderliggende measurement benchmark aanname. Deze veronderstelling houdt in dat primaire componenten van totale surveyfouten tegen de criteria meeteffecten zijn, terwijl selectie-effecten in alle gevallen klein en statistisch insignificant zijn. In het geval van een single-mode aan-huis benchmark, kan een aan-huis-interview na telefoon, schriftelijk of internet, de totale effecten (en meeteffecten) reduceren. Wanneer internet de gewenste mode is zal een aan-huis opvolging de totale effecten en meeteffecten doen toenemen, zodat een design waarbij alleen post en/of

internet wordt gebruikt, en niet een opvolging met aan-huis interviews wordt aanbevolen.

Hoofdstuk 6 sluit de dissertatie af met een brede discussie over de empirische bevindingen en relevante stappen in het reproduceren van vergelijkbare studies naar mode-effecten op surveyfouten. In het bijzonder worden de volgende conclusies getrokken:

1. Meet-effecten tussen interviewer en zelf in te vullen modes zijn de dominante componenten van totale mode-effecten in de VM. Afhankelijk van de measurement benchmark aanname is het óf beter om alleen interviewer modes, óf beter om alleen zelf in te vullen modes in mixed-mode designs van de VM te gebruiken. Een voordeel van zelf in te vullen modes zou kunnen zijn dat veel attitudevragen minder toevallige fouten hadden (hogere item betrouwbaarheid) dan in modes met een interviewer.
2. De single-mode designs in de VM laten alleen kleine verschillen zien in selectiefouten op sociaal-demografische variabelen, en maar heel kleine en meestal insignificante verschillen op doelvariablen. Een algehele vermindering van selectiefouten door een sequentieel mixed-mode follow-up design is niet in de empirische data gevonden, wat suggereert dat één van de klassieke motivaties achter mixed-mode surveys niet geld voor de VM. Gebruik maken van goedkope single-mode survey designs zoals internet of post zou daarom kunnen worden overwogen, maar moet dan wel worden geëvalueerd tegen het gewenste meetcriterium.
3. Het repliceren van de empirische bevindingen is zeer wenselijk. De afhankelijkheid van surveyfouten van onderwerpen, vragenlijsten en populaties, zou een grote invloed op de resultaten kunnen hebben.
4. Re-interview designs, zoals het within-subject design (WSD), zouden voor zeer waardevolle aanvullende gegevens voor het uitsplitsen van de overlappende (confounded) meet- en selectie-effecten in mixed-mode surveys kunnen zorgen.
5. De keuze voor meet- en selectie- benchmark-modes van voorkeur, is een belangrijke stap in mixed-mode analyse, om mode-effecten op systematische surveyfouten een inhoudelijke interpretatie te kunnen geven.

6. Een voordeel van analyses van toevallige en systematische foutvariantie is dat een benchmark-mode voor meten niet nodig is, maar zulke analyses houden geen rekening met bias en kunnen daarom zonder additionele informatie niet worden vertrouwd.
7. Additionele analyses van selectiefouten zouden benchmark-schattingen kunnen extrapoleren naar de populatie, en indicatoren van de representativiteit van surveys kunnen gebruiken om de keuze voor een selectie-benchmark te onderbouwen.
8. Experimentele gerandomiseerde mode vergelijkingen zijn belangrijke stappen in het ontwerpen van mixed-mode surveys, maar de conditionele single-mode metingen en selectie-effecten geobserveerd in deze designs zijn niet per se gelijk aan de effecten die worden geobserveerd in sequentiële of concurrerende mixed-mode surveys. Daarom moet de evaluatie van mode-effecten uiteindelijk mixed-mode data overwegen.

Samenvattend pleit dit proefschrift voor een sterkere methodologische aanpak voor mixed-mode onderzoek. Er wordt betoogd dat toekomstig onderzoek zich primair op twee aspecten moet richten. Ten eerste zou de statistische theorie van mixed-mode surveyfouten verder moeten worden ontwikkeld. Ten tweede wordt benadrukt dat een belangrijke volgende stap het ontwikkelen van methodologie voor het corrigeren van meeteffecten in mixed-mode data is. De re-interview aanpak ontwikkeld in dit proefschrift, zou waardevolle aanvullende data kunnen opleveren, die kan worden gebruikt in de ontwikkeling daarvan.