3

Internet Surveys as Part of a Mixed-Mode Design

Edith D. de Leeuw
Department of Methodology and Statistics
Utrecht University
Utrecht, the Netherlands

Joop J. Hox
Department of Methodology and Statistics
Utrecht University
Utrecht, the Netherlands

3.1 INTRODUCTION

In 1788, Sir John Sinclair conducted the first documented survey. Lacking funds for a full statistical census, Sinclair mailed out questionnaires to ministers of all parishes in the Church of Scotland with more than a hundred questions about their parish. He pursued them relentlessly using a mixed-mode strategy, with "statistical missionaries" to hurry up late responders, and follow-up letters of which the last was written in blood red to suggest with "the draconian colour of his ink" what would happen to the nonrespondents. It took 23 reminders, but Sinclair achieved a 100% response, and his study is thereby one of the first examples in which a mixed-mode strategy is highly successful in reducing nonresponse. By using a mixed-mode strategy for contact and reminders only (i.e., follow-up letters via mail and in-person "statistical missionaries"), but keeping the data collection itself restricted to one mode (i.e., the written accounts of Scottish ministers), Sinclair also made sure that mode effects would not threaten data integrity. The results were published between 1791 and 1799 in the Statistical Account of Scotland, and these accounts are still part

of the online National Data Centre at the University of Edinburgh. This account contains not only demographic statistics such as age distributions, life expectancies, and estimates of the total population, but also data on social statistics and lifestyles. Thanks to Sinclair, we now know that in the 18th century the women of Inveresk organized football matches between the married and unmarried women, and the former invariably won (Hacking, 1990; Heiser, 1996).

Surveys are part of the ever-changing cultural and technological context of society, and survey methodology changes over time. We no longer write to clergymen and use them as informants about their parishioners; instead, we survey people directly using today's technology. Still, common to all surveys—then and now—is that they require a two-way dialogue and that certain assumptions for the collection of valid data remain unchanged. These assumptions can be summarized in the four cornerstones of data quality: good coverage of the intended population, probability sampling, low nonresponse error, and accurate measurements (Groves, 1989; Biemer & Lyberg, 2003; De Leeuw, Hox, & Dillman, 2008a). Groves (1989) also added cost efficiency to the desiderata of good surveys. Like Sinclair, we usually lack the funds to do a census or even a large-scale interview survey, and one of the major attractions of online research is its low cost per completed questionnaire. But the Internet as data collection medium offers more advantages, such as the potential for using complex questionnaires and visual and auditory stimuli, and the quick turnaround time (e.g., De Leeuw, 2008; Chapter 1 in this volume).

While acknowledging the immense potential of Internet surveys, Couper (2000) pointed out that coverage error and nonresponse error are the biggest threats to inference from Internet surveys; see also Chapter 5 in this volume.

Although Internet coverage is growing, we get a very diverse picture when we look internationally. In his extensive inventory of Internet access at home across Europe, Blyth (2008a, 2008b) cites percentages ranging from 86% in the Netherlands and 83% in Sweden to 15% for Turkey and 21% for Bulgaria. These data are based on the Eurobarometer 2007; the Eurobarometer has been collecting data on access to technology EU-wide on a regular basis, using *personal interviews* with adults. For the United States, the percentage of households with Internet access is estimated at around 62% (see Chapter 1 in this volume). The differences in coverage across countries are so large that Blyth (2008a) argues that in order to

provide cost-effective international measurements, we must embrace a mixed-mode strategy. But even in countries with high Internet penetration, there may be a substantial risk of coverage bias because online access varies widely across demographic groups; thus a mixed-mode strategy may be called for. For instance, in the Netherlands—a country with coverage of over 80%—Internet access is unevenly distributed over the population, with highly educated, younger, and native Dutch people more often having Internet access (Bethlehem, 2008). Similar indications of a digital divide have been found in other countries; for instance, for the United States, Rookey, Hanway, and Dillman (2008) report that Internet users are younger, have higher incomes, and have more education; Couper, Kapteyn, Schonlau, and Winter (2007) find similar differences in socioeconomic status and age. That the digital divide can be substantial is shown by Couper (2008, p. 86) using data from the Pew Internet and American Life Project: while 91% of college graduates have Internet access, only 40% of those with no high school degree have access to the Net.

Besides noncoverage, nonresponse and its associated potential for nonresponse error are a threat to data quality (Couper & De Leeuw, 2003). Unfortunately, Internet surveys do not achieve high response rates, as recent meta-analyses show. In an early study, Cook, Heath, and Thompson (2000) report an average response rate of 34.6% (SD = 15.7%) based on 56 Internet surveys reported in 39 studies with complete data on (non)response. When Internet surveys are compared to other modes of data collection, they produce overall lower response rates than other modes. Shih and Fan (2008) summarized 39 comparisons of Internet and postal mail surveys that were published in the last 10 years. They found that the unweighted average response rate of paper mail surveys was about 10% higher than that of Internet surveys, which had an average response of 34%. This is in the same range as the results of Lozar Manfreda, Bosnjak, Berzelak, Haas, and Vehovar (2008), who compared Internet surveys to different modes of data collection (e.g., mail, telephone, face-to-face, fax). When analyzing 45 published and unpublished experimental comparisons, they found that Internet surveys yield on average an 11% lower response rate. There are several ways to improve the response rate of a single Internet survey, including incentives and reminders (for an overview, see Chapter 1 in this volume). Switching modes—that is, following up with nonrespondents using a different mode of data collection—can be very effective in increasing response rates to Internet surveys (Couper, 2008, p. 342). However,

the mode sequence should be planned by the researcher, going from most affordable to more costly methods; offering potential respondents a choice between methods does *not* increase the overall response (Dillman, Smyth, Christian, & O'Neill, 2008).

The above data all refer to response in cross-sectional (ad hoc) Internet surveys; no detailed overviews are available for response rates in Internet panels. A first attempt was made by Willems, Van Ossenbruggen, and Vonk (2006), who performed an extensive study of online panels in the Netherlands. They extensively analyzed the results of a comparative survey that was hosted by 19 commercial Dutch online panels. The questionnaire used was an omnibus, completion took on average 12.9 minutes, and no reminders were sent. The response varied between 18% and 77%, with an average response of 51%. Although response rates as high as 70% can be reached in an online panel, one should take into account that these data are panel data. Sikkel, Hox, and De Leeuw (2009) warn that response figures for panels may be misleading because the bulk of the nonresponse takes place during the initial recruitment phase; this initial nonresponse should be taken into account when reporting response rates.

In sum, two central problems of Internet surveys are undercoverage (due to limited Internet penetration) and low response rates, while major advantages are its affordability and cost-effectiveness. Mixed-mode designs involving Internet surveys offer an attractive alternative, providing an opportunity to compensate for the weaknesses of Internet surveys yet keeping costs affordable.

However, mixed-mode designs introduce a new problem, that of data integrity. Only when data collected with different modes produce the same results—that is, only when data equivalence is reached—is it allowed to combine these data into one data set. Therefore the main questions concerning all mixed-mode studies are: May data that are collected through different modes be combined in one cross-sectional study? May data collected with different modes at different time points be combined in a longitudinal analysis? May data that are collected through different modes be compared over studies or countries?

In this chapter we focus on Internet surveys as part of a mixed-mode design, and not on mixed-mode research in general. For the latter we refer to Dillman, Smyth, and Christian (2009), De Leeuw, Hox, and Dillman (2008b), De Leeuw (2005), and Roberts (2007). We start with some preliminary notes on data collection and mixed-mode designs. In Section 3.3 we

review empirical evidence regarding mode effects and Internet surveys. We then discuss optimal design for mixed-mode situations involving the Internet in Section 3.4, and end with recommendations for the design and analysis of mixed-mode surveys.

3.2 AVAILABLE DATA COLLECTION METHODS

3.2.1 Which Mode to Choose

In survey research there are two basic forms of data collection, self-administered questionnaires and standardized interviews, and these are mainly characterized by the absence or presence of an interviewer. But there are many variations possible. Interviews can be performed face-to-face or over the telephone, and self-administered questionnaires can be handed over by an interviewer during an interview or sent by paper mail or through the Internet. Furthermore, computer-assisted equivalents are available for both face-to-face interviews (computer-assisted personal interviewing, CAPI) and for telephone interviews (computer-assisted telephone interviewing, CATI). There are even self-administered computerized forms that are introduced by an interviewer; prime examples are computer-assisted self-interviewing (CASI), which can be used in a face-to-face situation, and interactive voice response (IVR), which is introduced over the telephone. These interviewer-introduced self-administered surveys, which are a form of mixed-mode surveys, are usually administered when very sensitive questions are being asked; the aim is to reduce social desirability bias.

Besides the presence and absence of an interviewer, data collection methods also differ on two other important dimensions: how the information is presented (visual, aural, or both) and how the respondents convey their answer (spoken, written, or typed); for a more detailed description, see De Leeuw (1992, 2008). These factors not only are of theoretical importance but also have a direct impact on data quality, because they influence the cognitive burden for the respondents, their sense of privacy, and the question–answer process as a whole (cf. Tourangeau, Rips, & Rasinski, 2000). These factors should be taken into account when designing optimal questionnaires for mixed-mode studies; we will come back to this in the section on questionnaire development.

Deciding which data collection mode is best in a certain situation is often complex and depends on many factors, of which the most important are the population under investigation, topic, types of questions to be asked, available time, and funds. When choosing a specific data collection mode, one wants to reduce the total survey error as much as possible, which means taking into account coverage and sampling error, expected nonresponse, and desired data quality. Each data collection method has its advantages and disadvantages (for a detailed overview, see De Leeuw, 2008). By combining different data collections methods in one mixed-mode survey it is possible to compensate for the disadvantages and exploit the advantages of each mode, and survey methodologists have been proposing mixed-mode surveys as the best of all possible modes for decades (Dillman et al., 2009).

There are many forms of mixed-mode surveys, and each serves a specific goal. De Leeuw (2005) presents a detailed typology and discusses the advantages and disadvantages of each design. One can discern four main groups of mixed-mode design.

3.2.1.1 Contacting by Different Modes

The first group of mixed-mode designs focuses on optimizing the *contact* with the respondent. In this form one or more modes are used to contact the respondent; response is stimulated by a different mode. As the actual data collection in these cases is in one mode only, these mixed-mode designs have no negative implication for measurement error or data integrity, but will reduce nonresponse error or coverage error—a win-win situation. Examples are advance notification letters to establish legitimacy or send incentives before telephone interviews and Internet surveys, screening or selecting respondents by telephone while the actual data collection is done through the Internet, and reminders in a mode different from previous contacts. Sir John Sinclair did in fact use this type of mixed-mode design as early as the 1790s, when he decided to send "statistical missionaries" in person to hurry up the overdue written accounts of Scottish ministers.

3.2.1.2 Another Mode for Specific Questions in the Questionnaire

In this type of mixed-mode design, a more private, second data collection method is used to collect data from all respondents during a single

data collection period for a specific *subsection* of the questionnaire. The motivations for choosing this design are to reduce social desirability bias for sensitive questions and to reduce overall measurement error. Usually a mix of interview and self-administered forms is used to exploit the strong points of both methods. For instance, within an interview a self-administered form of data collection such as CASI or Audio-CASI is used for sensitive questions to reduce social desirability and enhance privacy, as neither the interviewer nor any other person in the vicinity will know the answers given; all other questions (e.g., complex questions, household composition) are administered by the interviewer, who may provide assistance when necessary. This situation has only positive points and is not a cause for concern regarding data integrity.

3.2.1.3 Different Modes for Different Respondents

A different situation arises when one mode of data collection is used for some respondents of a sample and another mode for others in that same sample in order to collect the *same* data. In other words, different modes are used for different respondents for all questions in one survey. This is usually done in order to reduce coverage and nonresponse errors. For instance, to reduce undercoverage of special groups an Internet survey is implemented together with a mail survey of all sampled units that do not have an Internet connection. Another application of this type of mixed-mode design is to offer multiple data collection methods in a specific sequence during a survey in order to reduce overall nonresponse. For instance, a telephone interview is implemented among the nonrespondents to an Internet survey. A third application is when different modes are used in different regions or countries—for instance, an Internet survey in countries with high Internet penetration and a telephone survey in countries with high telephone coverage, while a face-to-face interview is conducted in countries that have neither. The motivation for this type of mixed-mode design is to respond to cultural differences and variations in survey capabilities in different countries and keep the overall survey costs manageable. In all these situations, data are collected using a different data collection method for discernable (sub)groups of respondents and these data are then combined and analyzed in one data set. Here the question of data integrity does play a role, and differences between subgroups may be confounded with mode

differences. For example, do the nonrespondents to an Internet survey really differ in opinions, or is this a (telephone) mode effect? In the former the mixed-mode strategy helps to reduce nonresponse bias, while in the latter the mixed-mode strategy adds bias through differential measurement error.

3.2.1.4 Alternating Modes in a Longitudinal Design

The fourth group of mixed-mode designs involves alternating modes over time in a longitudinal study or a panel. Respondents are surveyed at different time points using different modes. Here practical considerations, such as the availability of a good sampling frame, and costs are the main reasons for this approach. Sometimes addresses are available, but e-mail addresses are not and have to be collected first; sometimes no sampling frame is available and area probability sampling using face-to-face interviews is the only option. This flexibility together with the greater likelihood that an interviewer will gain cooperation in person at the doorstep and the better opportunities for screening make the face-to-face survey a favorite for the baseline study in a panel. To reduce costs a more efficient and less expensive method (e.g., an Internet survey) may be used after the first wave. However, when modes are alternated in a longitudinal design, time and mode effects are confounded, as the change in data collection mode may introduce differences in the measurements, and it is difficult to decide if a change over time is a real change or the result of a change of mode.

In sum, there is no easy solution for the problems mentioned in the last two types of mixed-mode designs. Depending on the survey situation, one has to decide upon the optimal design while carefully appraising the different sources of error. Only after careful consideration can one decide if the expected mode effects are serious enough to avoid mixed-mode designs or if the advantages of mixing modes outweigh the risks.

Using different modes for different parts of a sample or alternating modes in a longitudinal survey can introduce measurement error because people may respond differently to the same questions depending on whether the question is posed through the Internet or with another data collection mode. How serious is this threat? In the following section we provide an overview of empirical mode comparisons with regard to potential mode differences involving Internet surveys.

3.3 A REVIEW OF EMPIRICAL EVIDENCE OF MODE EQUIVALENCE

3.3.1 Threats to Data Integrity

The goal of mixed-mode surveys is to combine data from different sources into one data matrix for analysis. This assumes that data collected through different modes are equivalent, but this assumption is not necessarily true. Data from different sources may differ for the following reasons: (1) because different modes may lead to a different sample composition, (2) because different question formats are employed in different modes, and (3) because the modes themselves lead to different response processes (e.g., De Leeuw, 2005; Dillman & Christian, 2005).

The fact that respondents to different modes may have different background characteristics and therefore provide different answers (point 1) is an advantage, not a disadvantage, of mixed-mode surveys. This is not a threat to our data, but something we aim at and want to achieve. We should remember that one of the main reasons for mixing modes is to overcome coverage and nonresponse errors of single Internet surveys, which means that we explicitly do want to bring in different groups using different modes.

When data from different sources are combined, the danger always exists that one and the same concept is measured using questions that are differently worded or even have a different question format (point 2). As data collection methods have different philosophies of question writing and question format (Dillman, 2008), these different approaches to questionnaire construction may unintentionally lead to nonequivalent questionnaires when an Internet survey is mixed with other data collection methods. A prime example is the way a list of response options is offered. Because of the richness of the visual channel in Internet surveys, a list of response options is visually displayed on the screen (e.g., strongly agree, agree, somewhat agree, neutral, somewhat disagree, disagree, strongly disagree), a procedure akin to mail surveys and interviewer show cards in face-to-face interviews. In telephone surveys branching (unfolding) is used in these cases—for example, asking first the direction of an opinion (e.g., agree, neutral, disagree) and following that up with the intensity (e.g., is this strongly agree, agree, or somewhat agree). These

different formats affect the responses, leading to differences between modes that offer the full list of response options visually (Internet, mail) and unfolding in a telephone interview. From past research we know that question format does have an effect on the responses and the response distribution even within a single mode (e.g., Sudman & Bradburn, 1974; Schuman & Presser, 1981; Christian, Dillman, & Smyth, 2008); therefore, question-format effects may be one of the main causes for mode effects in standard mixed-mode designs. To avoid unwanted divergence across modes one should avoid differential questionnaire construction and aim at equivalent questionnaires when employing a mixed-mode design.

The data collection mode itself can influence the data (point 3). Modes vary in terms of (1) interviewer versus self-administered questionnaires and the associated interviewer effects; (2) in the way information is transmitted, the survey question is posed, and the answer is recorded (e.g., aurally versus visually, spoken versus written versus typed); and (3) in general media-related factors, such as knowledge, experience, and social customs related to the medium. These mode characteristics influence the potential for social desirability bias, the difficulty of the task for the respondent, respondent motivation, and the question-answer process in general, which in turn influence data quality. For an overview, see De Leeuw (1992, Chapter 2); Roberts (2007); and Tourangeau et al. (2000, Chapter 10).

3.3.2 Review of Mode Differences for Traditional Data Collection Methods

The influence of data collection method on data quality has been extensively studied for the traditional data collection methods, that is, face-to-face interviews, telephone surveys, and self-administered paper mail questionnaires. These older reviews can provide some insight into media-related effects and potential mode effects involving mixes with Internet surveys. De Leeuw (1992, Chapter 3) performed a meta-analysis of 67 articles and papers reporting mode comparisons. The resulting overview showed consistent but usually small differences between methods, suggesting a dichotomy of survey modes in those with and without an interviewer. Especially with more sensitive questions, self-administered surveys performed better with less social desirability bias in answers and more reporting of sensitive behaviors such as drinking. This is promising for Internet surveys, which are a form of self-administered surveys.

A limited number of studies investigated specific response effects, such as recency and primacy effects. In a visual format, such as a self-administered questionnaire, respondents think in the order in which the response categories are presented and are more likely to choose those presented at the beginning of a list of response alternatives than those at the end (a primacy effect), while in an auditory format, respondents are expected to wait till the interviewer has read the whole question and are more likely to start thinking about the last alternatives read to them (a recency effect). The evidence on this is mixed. Dillman et al. (1995) found inconsistent evidence for primacy effects in mail and recency effects in telephone surveys in a large number of experiments. These inconsistent findings may be due to interaction effects with social desirability. In general, mail surveys produce fewer socially desirable answers than telephone surveys, and when the last response category is also the less socially desirable answer, this may counteract the primacy/recency effects. This may have implications for mixing self-administered modes, such as Internet surveys with interviewer-administered modes.

Finally, in a carefully designed experiment, De Leeuw (1992, Chapter 6) investigated reliability and consistency of answers. Again the main difference was between self-administered and interviewer-administered surveys. In the self-administered mail survey, where the respondent is in control and can read the questions and answer at his or her own pace, more consistent answers were given and less random error was detected in the answers. Again, this is promising for Internet surveys.

3.3.3 Measurement Error in Internet Surveys Compared to Other Data Collection Methods

The above review was limited to comparing paper-and-pencil self-administered questionnaires with face-to-face and telephone interviews; however, the results are of importance for Internet surveys as well. Although Internet surveys are a new form of data collection, one should remember that Internet surveys are a form of self-administered questionnaires and many of the benefits of self-administration should apply, such as absence of interviewer effects, lower social desirability bias, visual presentation, and ability for the respondent to set the pace. On the other hand, there are also differences between paper self-administered questionnaires and computerized forms. Internet surveys may be accessed from any

place (e.g., home, office, library, Internet café) at any time, and Internet surveys are more interactive and the medium itself is clearly different from paper and pencil; for instance, the Internet more easily allows for multitasking, scanning the page and quickly skipping from one topic to the next, and satisficing (cf. Krug, 2006), all of which may influence data quality. These considerations have led to a new series of empirical mode comparisons, and we will first review comparisons with paper-and-pencil self-administered forms and then comparisons with either telephone or face-to-face interviews.

3.3.3.1 Self-Administered Questionnaires and the Internet

The importance of the medium of administration for data quality has long been recognized in diagnosing and assessment. When computerized forms of tests were introduced, the American Psychological Association (1986; p. 18) explicitly stated that "the equivalence of scores from computerized versions should be established and documented before using norms of cutting scores obtained from conventional tests." This led to numerous empirical comparisons between computerized and paper-and-pencil versions of well-known tests, which in turn resulted in quantitative summaries and meta-analyses. In general, the mode of administration had no statistically significant effect. In one of the first meta-analyses, Bergstrom (1992) found negligible differences between paper-and-pencil and computerized tests for general aptitude based on 15 studies of adults and high school students. These results were confirmed by Mead and Drasgow (1993) with a notable exception. In a meta-analysis of 29 studies comparing paper and computerized tests for cognitive abilities among young adults and adults, they found that power tests (that is, ability tests without a time limit) were highly equivalent, with a cross-mode correlation of 0.97, but speed tests (tests measuring cognitive processing speed, where simple tasks have to be processed within a time limit) were clearly less equivalent, with a cross-mode correlation of 0.72. Mead and Drasgow (1993) interpret the mode effect for speed tests as an effect of the importance of perceptual and motor skills in responding quickly to time-pressured tests. These overviews go back to comparisons as early as the late 1970s, when far fewer people were acquainted with and were using computers, and we may expect that the present "Nintendo generation" is better trained in the perceptual and motor skills needed for quick and accurate reactions to a

computer-offered stimulus. Indeed, later meta-analyses (Kim, 1999; Wang, Jiao, Young, Brooks, & Olson, 2007, 2008) confirm that for high school students computer-assisted and paper achievement tests are equivalent. However, it seems wise to allow ample time when less computer-literate groups, (e.g., the elderly) are studied (De Leeuw, Hox, & Kef, 2003).

Also, for noncognitive instruments, which ask for more subjective information, equivalence has been established. Gwaltney, Shields, and Shiffman (2008) performed a meta-analysis of 65 studies comparing electronic and paper-and-pencil self-reported patient outcome measures on such diverse topics as health status, anxiety, depression, pain, quality of life, and mood. Their results show that computerized and paper-and-pencil measures produce equivalent scores: The mean differences were very small and not significant, and correlations across modes were very high and similar to correlations between repeated administrations of the same paper-and-pencil measurement. Finally, Richman, Kiesler, Weisband, and Drasgow (1999) investigated social desirability distortions; their meta-analysis reports that in the case of computerized versus self-administered paper questionnaires a near zero overall effect was found, both for direct measures of social desirability distortion and for social desirability distortion inferred from other scales. However, a very interesting moderator variable was identified: When respondents were not assured of anonymity, were identified, or were in the close presence of others, they were less willing to reveal personal weaknesses in the computerized form than with paper and pencil. Although the effects were small, this may be of concern when very sensitive data are collected through the Internet, and it underscores the importance of confidentiality assurances in Internet surveys.

For Internet surveys this evidence of test equivalence is promising indeed: After more than three decades of investigation, computerized tests appear to be accepted as being valid and reliable alternatives to traditional methods (Epstein & Klinkenberg, 2001), and online tests can be seen as a special case of computer-assisted testing. Still, there are differences between computerized and Internet administration. Computerized testing is usually done under very controlled conditions, while an Internet survey or online test may be completed from any number of locations at any time and relatively free of controls. Despite these differences, test data collected over the Internet and via paper and pencil appear to be largely equivalent, as Preckel and Thieman (2001) demonstrated for a new intelligence test. No differences were found for reliability and validity for the two versions; the

only difference found was in mean score, with online respondents scoring higher than paper-and-pencil respondents, which could be attributed to self-selection. Also, for noncognitive instruments, measurement equivalence of online and paper-and-pencil questionnaires could be established. In a large-scale multinational test in 50 countries, Cole, Bedeian, and Feild (2006) stringently tested equivalence for a leadership test consisting of 20 items. Not only did they find similar reliability coefficients and intercorrelations across modes, but they were also able to establish full equivalence across modes using multigroup confirmatory factor analysis. The only difference found, a slightly higher score in the Internet condition, could be attributed to differences in sample composition. Similar results for a range of personality measures were found by Meade, Michels, and Lautenschlager (2007) and by Ferrando and Lorenzo-Seva (2005). Meade et al. (2007) also reported that while they were able to establish equivalence of measures across modes in a strict experiment where respondents were allocated at random to the Internet or to paper and pencil, the results were more complicated when respondents were given a choice between modes. They comment that although measurement equivalence exists for format (Internet versus paper) when controlling for choice, it may not exist for people allowed to choose and those not allowed to choose, even if respondents (as was the case here) were of the same age and level of education. This could have implications for a well-known form of mixed-mode surveys where respondents are offered a choice of modes. Not only is there evidence that a mode choice does not raise the response rate and may even lower it (see also Dillman et al., 2008), the study by Meade et al. (2007) suggests that it also may offer a potential threat to data integrity.

In sum, the results are promising for mixing Internet and other forms of self-administered questionnaires. Generally, it seems reasonable to assume that respondents use the same psychological processes and metric when responding to Internet and other forms of self-administered questionnaires. However, most studies reviewed are strict migrations, where the exact text of the paper instrument was ported to the computer screen without making substantive changes in the content. When substantial changes in the questionnaire are made or where layout changes substantially, affecting users' perception and ability to respond (e.g., scrolling, drop-down boxes), equivalence is not guaranteed and new studies will be necessary.

3.3.3.2 Interviews and the Internet

There are fewer comparisons with interview surveys, either telephone or face-to-face, and as a consequence there are as yet no comprehensive meta-analyses summarizing mode effects for Internet versus interview surveys. However, one effect is consistently found in the available studies: Internet surveys appear to give rise to less social desirability bias than interviews. In this sense, Internet surveys are indeed more like self-administered questionnaires and share their benefits, as Couper (2008) postulated.

Self-administered questionnaires have been found again and again to lead to less social desirability bias and more openness in answering sensitive questions when compared to interview surveys. This was the case with paper-and-pencil questionnaires and mail surveys (for a meta-analysis, see De Leeuw, 1992) and also with computer-assisted self-administered questionnaires (see, for instance, Tourangeau and Smith, 1996, and Turner et al., 1998).

These results have now been replicated for Internet surveys. In a controlled experiment in Belgium, Heerwegh, Billiet, and Loosveldt (2005) found more socially desirable answers on questions regarding rights for immigrants in a face-to-face interview than in a Web questionnaire. Bronner and Kuijlen (2007), when comparing face-to-face (CAPI), telephone (CATI), and Internet interviewing (which they label CASI@home), also found more reporting of socially undesirable behavior, such as violations of the law, in the Internet condition than in the CAPI or CATI condition. As was the case with Heerwegh et al. (2005), an experimental design was used, and the differences could not be attributed to self-selection or differential respondent characteristics.

In the United States, Link and Mokdad (2005) found more self-reported heavy drinkers in an Internet survey compared to those in a telephone interview. This result remained strong and significant after adjusting for different demographic characteristics of respondents in both modes. Similar results were found in the Netherlands, where less drinking of alcoholic beverages and more ecologically friendly behavior was reported over the phone compared with the Web for the same population (Van Ewijk, 2004).

Krauter, Presser, and Tourangeau (2008) compared CATI, interactive voice response (IVR), and Internet surveys in an experimental setting and confirmed and extended these findings. Internet administration increased reporting of sensitive information among alumni in the United States, such as more yes answers to, and less skipping of, questions asking for undesirable or sensitive information (e.g., dropping a class, GPA). Krauter

et al. (2008) also had access to record data and found a higher accuracy in Internet surveys; they report that Internet surveys increased both the level of reporting sensitive information and the accuracy compared to telephone (conventional CATI), with the more private self-administered telephone survey (IVR) in between. Finally, Dillman et al. (forthcoming) compared mail, telephone, IVR, and the Internet and found that more extreme positive answers on satisfaction–dissatisfaction questions were given in the telephone and IVR conditions than when using paper mail or Internet surveys; a careful check showed that this result could not be accounted for by a tendency toward recency over the phone. The less positive answers in the self-administered forms very well could be attributed to more openness and less social desirability bias. Dillman et al. (forthcoming) point out that the visual versus aural communication channel could play a role here too, and warn that mixing modes that depend upon different communication channels (i.e., visual versus aural), may introduce measurement differences that cannot be ignored.

Although the results for social desirability are clear—more openness in Internet than in interview surveys—the pattern is far less clear for other indicators of data quality. Regarding item nonresponse, contradictory findings are reported. Heerwegh and Loosveldt (2008) in a Belgium experiment report more "do not know" answers and more item nonresponse in an Internet survey than in a face-to-face interview of students on attitudes toward immigrants. Van Ewijk (2004) also reports higher item nonresponse for Internet than for telephone interviews in a Dutch experiment using an omnibus questionnaire. Van Ewijk attributes the lower nonresponse in interviews to the fact that it is easier to skip a question in a self-administered questionnaire than to say "do not know" to a real person. However, Fricker, Galesic, Tourangeau, and Yan (2005) report less item nonresponse in an Internet survey than in a telephone interview for a study on attitudes and knowledge toward science. This could be explained by the fact that in the Internet condition respondents were prompted, whereas in the telephone condition "no opinion" was accepted without further probing. To make matters even more complicated, Link and Mokdad (2005) report a very low level of item nonresponse in both Internet and telephone surveys and a slightly higher level in a mail survey on drinking. A similar result is reported by Oosterveld and Willems (2003), with no differences in item nonresponse between Internet and telephone in a survey on finances. In the last two studies the questions asked were of a highly sensitive nature.

Much depends on the nature of the questions and the way the survey is implemented, as Toepoel (2008, Chapter 2) argues in discussing her findings that putting more items on the screen increased item nonresponse; more experimental research is clearly needed here.

It has been argued that Internet surveys may give rise to more satisficing than surveys in which respondents are interviewed (e.g., Krug, 2006; Couper, 2008). This tendency to satisfice may account for more missing data. It may also introduce other response effects. For instance, Heerwegh and Loosveldt (2008) report that Internet respondents differentiate less on rating scales than respondents in a face-to-face interview. Fricker et al. (2005) found similar results when comparing Internet and telephone interviews. However, Smyth, Christian, and Dillman (2008) and Van Ewijk (2004) show that check-all-that-apply questions, which could be seen as encouraging satisficing behavior, perform less well and endorse fewer response options than yes/no questions in *both* Internet surveys and telephone interviews, which suggests a question-form effect instead of a mode effect. Furthermore, Oosterveld and Willems (2003) report more answers to open questions for Internet than CATI when comparing these two conditions. Similar findings are reported by Fricker et al. (2005), who report that Internet respondents performed better on knowledge questions than telephone respondents did and took longer to complete especially the open questions, which points to more thorough cognitive processing and less satisficing for the Internet.

In sum, Internet surveys perform better than both face-to-face and telephone interviews when sensitive questions are asked. Evidence on satisficing behavior is mixed, and there are studies that clearly show that respondents in Internet surveys provide more answers to open questions than in telephone interviews. Finally, some studies report more item nonresponse in Internet surveys, while other report less or no differences at all. Again, how questionnaires were designed and how the actual survey was implemented may be crucial for the quality of answers.

3.4 CONSEQUENCES OF MIXED-MODE DESIGN FOR QUESTIONNAIRE DEVELOPMENT

The empirical mode comparisons cited above show relatively small differences between Internet and other modes of data collection, with the

exception of Internet and interview mixes for sensitive questions. This seems reassuring, but usually in experimental mode comparisons extreme care is taken in designing and implementing equivalent questionnaires. In daily survey practice, differences in question wording and question format between specific modes may be the biggest threat to data integrity in mixed-mode surveys, as each survey mode has different conventions in developing questionnaires (Dillman, 2008). Examples are offering a "do not know" category to the response scale in self-administered questionnaires and withholding this in an interviewer-administered mode, using an open question in a telephone interview and a closed question with multiple response categories in an Internet survey, or shortening or changing the structure (e.g., branching or unfolding) of a question with a long list of response categories for telephone use. Implementing standard questionnaires, which are designed separately for individual modes, enhances unwanted measurement error in mixed-mode designs; to avoid this, one should explicitly design a special questionnaire for a mixed-mode survey.

3.4.1 Unwanted Consequences of Independently Designed Questionnaires

What would happen if in a mixed-mode design each mode is developed separately? Advocates of this approach argue that if each mode is optimized separately, this will reduce the total error in the combined data set, as some optimized modes have more error than others (e.g., more item nonresponse in a self-administered questionnaire) and in the combination the overall error is acceptable. The reasoning is that if one mode has strong properties, using these will reduce the error in the data resulting from that mode and thus reduce the total error. This is true only if the very strong assumption holds that all measurement errors associated with the mode are random error, because then we have one method with more random error and another with less. In the case of systematic error or bias (e.g., social desirability, acquiescence), it can be dangerous to optimize each method separately, as the bias may add up and result in increased overall bias for the combined data. As always, the burden of proof is on the researcher to demonstrate that the chosen design indeed results in better quality and that there is no added bias, for instance by embedding a small mode experiment.

3.4.2 Robust Questionnaire Design for Mixed-Mode Surveys: Unified Mode Design

When developing questionnaires for a mixed-mode approach, one should focus on the goal of the mixed-mode design and on the reduction of mode effects by measurement error. It is important to analyze the different modes, recognizing when the media differ and listing the limitations and extra features of each method. When mixing two modes that both use the visual channel—for instance, paper self-administered and Internet surveys—one can design questionnaires using this visual channel and offer longer lists and visual stimuli. When mixing telephone interviews (aural mode) and Internet (visual mode), one cannot use long lists and must find other solutions, such as using an open-ended question or using an unfolding procedure, which can be programmed in both modes.

To provide researchers with general rules in designing questionnaires suitable for modes, Dillman (2007) proposed a unified mode design, or unimode design: designing questions and questionnaires to provide the same stimulus in all survey modes in order to reduce differences in the way respondents answer survey questions in different modes. Dillman (2007, pp. 232–240) outlines several principles to construct unified mode questionnaires. These include making response options the same across modes, incorporating response options in the stem of the question, and using the same descriptive labels for response categories. A good example of a unimode design for the short form of the U.S. census and its complexities is given by Martin et al. (2007), who used the central design principle that all respondents should be presented with the same question and response categories, independent of mode.

Unified mode design is often viewed as aiming at the lowest common denominator. That is not necessarily true; one should see it as an attempt to design robust questionnaires. A good example is check-all-that-apply questions, which are often used in Internet surveys, versus a series of yes/no questions, which is often used in telephone interviews. Smyth et al. (2008) showed that in Internet surveys a series of yes/no questions also performs better than the traditional check-all-that-apply questions. Couper (2008) advises to use check-all-that-apply questions sparingly, but recommends avoiding very long lists of yes/no questions, as this may increase the risk of break-offs. Usually when designing a mixed-mode telephone-Internet

survey one would choose a reasonable list of questions to avoid break-offs in telephone and Web; in that case using a uniform yes/no format is no problem and increases the quality in both modes.

3.4.3 Beyond Unified Mode Design: Cognitive Equivalence of Questionnaires

Finally, one can go beyond designs that force different modes to use exactly the same questions. When questions are considered as stimuli that initiate a response process in the respondents, the perspective changes from offered stimulus to perceived stimulus. Using the same stimulus in different modes does not guarantee that the same question–answer process will be initiated, nor that respondents in one mode will perceive the question in the same way as respondents in a second mode. For example, a question in a telephone survey is not necessarily the same perceived stimulus as that same question when it is posed in an Internet survey, since the visual mode may change the meaning of the question and may therefore present a different *perceived* stimulus to the respondent than the aural (telephone) mode (cf. Tourangeau et al., 2000). Thus, in designing questions for a mixed-mode study, one should go a step further and aim at achieving cognitive equivalence of the perceived stimulus, rather than literal uniformity of questions across modes. This may imply that a slightly different question format for each mode is necessary to achieve the needed cognitive equivalence. De Leeuw (2005) labeled this "generalized mode design," while Dillman et al. (2009) use the term "mode-specific design" for the same concept and Couper (2008) emphasizes comparability of data. Whatever the term used, a prerequisite for successful mixed-mode design is that the question designer must understand how differences between modes affect response. A good illustration of this is the work of Wine, Cominole, Heuer, and Riccobono (2006), who used specially designed pop-ups after a "no answer" in an Internet survey, to emulate the probes used in previous telephone interviews.

A study by Christian, Dillman, and Smyth (2005) provides some insight into how and why different question formats across survey modes lead to equivalent results. In a telephone interview that asked, "When did you start attending Washington State University?" only 13% of the respondents reported the month and year, as desired. Instead, most respondents gave comments such as "last spring semester," "fall 2002,"

or "this is my first semester." These responses were followed up by the interviewer to obtain the desired response format, showing the strength of interviewer-assisted surveys. In the initial Internet survey, where the response had to be provided in open text boxes, only 45% of the respondents answered in the required format (two digits for month and four digits for year). By decreasing the size of the month box relative to the year box, replacing "month" and "year" with the more precise language of symbols (mm/yyyy), and placing those symbols in natural reading order ahead of the appropriate response boxes, the percentage of people responding in the desired way increased from 45% to 95%. These results clearly illustrate how different wording approaches of the question (telephone and Internet survey) can lead to the same result, but through different mechanisms. In the telephone survey, the interviewer served as an intelligent system that could ask for more information and convert the answer to the desired format required by the CATI system. In the Internet survey, the emphasis was on answer space labeling and layout in order to get the respondents to respond in the desired format and avoid error messages. Thus, different wording produced the same results. Using the same wording, "What month and year did you begin the studies?" for both Internet and telephone in this case actually *decreased* the equivalence of the recorded answers.

Just as in comparative research, in an optimal mixed-mode design the burden is on the researcher to demonstrate that these different questions do indeed elicit equivalent responses. This requires that at least some other, correlated questions are kept identical across different modes, which can then be used to test the equivalence of questions in the questionnaire. This is similar to the strategy used to statistically adjust responses in different modes to make them equivalent.

3.5 CONCLUSIONS

3.5.1 Internet Surveys and Mixed Mode

From a total survey error perspective, mixed-mode designs for Internet surveys are very attractive: Mixing modes greatly increases coverage and leads to less nonresponse at affordable costs. But every coin has two

sides, and mixing Internet with other modes may lead to problems of data integrity.

A key methodological assumption in all mixed-mode surveys is that data from different modes can be meaningfully combined and compared—in other words, that there is measurement equivalence across modes. For Internet and paper-and-pencil self-administered questionnaires, measurement equivalence has been established in numerous cases, and when differences were found, these could be attributed to differences in sample composition or to self-selection. For Internet versus interview surveys, the situation is less clear. The results summarized in this chapter give confidence in well-designed mixes of visual self-administered modes, such as Internet and paper mail surveys. But mixing data collection modes that depend upon different communication channels (i.e., visual versus aural) may introduce nonignorable differences into the resulting data, and one should be more careful with Internet surveys mixed with interviews. Still, in some well-designed experiments measurement equivalence was established for CATI and Internet surveys (e.g., Oosterveld & Willems, 2003).

One also has to keep in mind that in most empirical comparisons reviewed, the focus was on methodological research, and extreme care was taken in designing the survey and attaining equivalent questionnaires. When different question formats are used in different modes (e.g., two-step unfolding versus full scale) or where layout changes substantially alter the response process (e.g., scrolling), equivalence of measurement is not guaranteed and measurement differences between modes are only to be expected.

3.5.2 Designing for a Mixed-Mode Approach Including Internet Surveys

In the design phase, using a questionnaire design strategy that encourages comparability of questionnaires reduces the impact of mode differences. For instance, the forced-choice yes/no format instead of a check-all-that-apply question format helps respondents to carefully process the questions, as is demonstrated by Smyth, Dillman, Christian, and Stern (2006), who also show that a forced-choice question not only produced more endorsed items than check-all-that-apply questions, it also took respondents longer to answer the forced-choice questions, suggesting less satisficing. Encouraging respondents to process the task carefully, take their time in answering, and allowing for ample time to complete a questionnaire

may be crucial issues for quality in Internet surveys, as was also pointed out by Mead and Drasgow (1993) when explaining the lack of equivalence between computerized and paper-and-pencil speed tests.

Other factors also play a role, such as keeping respondents motivated and making the survey a pleasant experience. This is illustrated by the findings of Van Meurs, Van Ossenbruggen, and Nekkers (2009), who developed a system to identify respondents with satisficing response patterns in order to flag dubious or fraudulent respondents in a Dutch online panel. There was no systematic pattern in the demographic profile of panel members who were flagged and those who were not, but the researchers found a systematic pattern related to the questionnaire involved. Questionnaires that were evaluated as less interesting, boring, or too long evoked more dubious responses. Van Meurs et al. (2009) conclude that the most effective way to prevent fraudulent responses is to improve the quality of the questionnaire.

The visual design of Internet surveys is important for the quality of the questionnaire, as Toepoel (2008) shows with her findings that putting more items on a screen not only increased item nonresponse but also led to less positive assessment of the questionnaire itself. For more information on the importance of visual design, see Chapter 7 in this volume. Furthermore, the interactive power of the Internet may be used to replace some of the interviewer's tasks and keep respondents motivated in order to reduce item nonresponse (see Wine et al., 2006) and stimulate responses to open questions (see Chapter 9 in this volume).

When investigating sensitive topics through the Internet, special care should be taken. In general, one should be careful to avoid mixing self-administered modes, such as Internet surveys, with interview modes when sensitive questions are being asked. Mixing Internet with other self-administered forms, such as postal surveys, does not pose problems in this regard. In general, self-administered survey modes, including Internet, give rise to more self-disclosure and less social desirability bias on sensitive topics. But respondent trust is of the utmost importance, and when respondents are not assured of confidentiality or anonymity they are less willing to reveal personal or sensitive information (e.g., Richman et al., 1999). Protecting the confidentiality of the responses is one of the basic ethical rules of the survey industry (see Chapter 6 in this volume), and in Internet surveys with sensitive questions one should make sure to convey this to the respondents. A researcher may follow all the privacy and security guidelines, but if respondents are not aware of this, they may

be inhibited when answering sensitive questions. Therefore, researchers should enhance the *perceived* security, especially as potential respondents may be wary of the security of the Web in general. There are several ways the perceived security may be heightened, such as by explicitly stating that answers are confidential, and by implementing a secure Web exchange with encryption during the session and making this clear to the respondent by a well-known icon such as a lock. Establishing trust is most difficult in "cold" Internet surveys, and it may be necessary to use another mode of communication, such as a paper advance letter, to assure the potential respondents of the legitimacy of the survey. When there is an existing relationship with the respondents, for instance in a good organized Internet panel, there is already a basic trust between researcher and respondent to build upon.

3.5.3 Embedded Experiments and Adjustment

3.5.3.1 Measurement Equivalence

Finally, in the analysis stage, researchers need to check equivalence of measurement across modes. This is especially important when different question formats or mode-specific optimization is used.

If the survey contains multi-item scales, multigroup structural equation modeling is a suitable tool to investigate measurement equivalence across modes. Measurement equivalence requires that a confirmative factor model fit the multi-item scale, and that the factor loadings and intercepts can be constrained equally across the groups. Some amount of difference is allowed; see Vandenburg and Lance (2000) for a review and discussion of the issues involved. The problem is that when measurement equivalence is not achieved, it is not clear whether this is the result of a mode effect or of a different sample composition in the two groups.

This can be disentangled to some degree by using subgrouping or propensity score methods. In the propensity score approach, logistic regression is used to predict membership of a specific mode sample, using available background variables. The propensity score is the predicted group membership. It can be used as a covariate, or the inverse of the propensity score can be used as a correction weight. For a review of using propensity scores in surveys, see Lee (2006). If measurement equivalence is achieved after propensity score weighting, we can ignore the difference in mode provided this weighting is used in the subsequent analyses.

If a difference remains, we can attempt to equate the scale scores across modes, as Van Buuren, Eyres, Tennant, and Hopman-Rock (2005) propose. Their method, termed "response conversion," attempts to transform responses obtained on different questions in different surveys onto a common scale. If this can be done successfully, meaningful comparisons can be made using this common scale. The first step in response conversion is the construction of a conversion key using a statistical model. In Van Buuren et al. (2005), the polytomous Rasch model is used, but a confirmatory factor analysis with strong measurement equivalence also would do. A prerequisite for response conversion is that there be sufficient overlap between the different items—in other words, that for some items there must be strong measurement equivalence. For a detailed methodological review, see Van Buuren et al. (2005).

3.5.3.2 Embedded Experiments

Different respondent selection in different modes is a difficult problem, and is hard to solve in daily survey practice. Also, alternating modes in a longitudinal design causes problems by a potential confounding of time and mode effects. A sound approach in all cases is implementing a small mode experiment including randomized assignment of respondents to the different modes in the survey in order to assess and compensate for potential mode effects (De Leeuw, 2005). This will increase the effort and costs, but especially in large national surveys, cross-national studies, and longitudinal surveys this is well worth it, and researchers are advised to allocate a small part of their budget for mode-effect investigations, just as part of the budget is often allocated to studies into nonresponse bias and adjustment.

A small mode comparison embedded into the data collection procedure enables the researcher to estimate potential mode effects, and if necessary to statistically adjust for it, using calibration, propensity score adjustment, or response conversion (e.g., Lee, 2006; Lundstrom & Sarndal, 2002; Van Buuren et al., 2005). For instance, in a longitudinal design a random subsample of respondents can be investigated using the initial data collection mode of the previous wave and the majority of the sample can be surveyed using the main mode of the wave (e.g., a small subsample is interviewed face-to-face, while the intended mode is Internet). If for practical consideration a random subsample is not possible, one could implement a small embedded mode experiment among those with Internet access,

where a random half of the subsample is surveyed using the Internet and the remainder using the telephone. Even if in this case the experiment is not performed on a random subsample of the whole population, it will still provide valuable information to extrapolate and assess the risk of mode effects (De Leeuw, 2005).

A good example is the experiment by Jäckle, Roberts, and Lynn (2007), who were able to disentangle the effects of interviewers and use of show cards in telephone and face-to-face interviews, using a subset of the core questions of the European Social Survey. The experiment took place in Budapest and the immediate surrounding area, where telephone penetration was high. Respondents were randomly assigned to three groups: (1) face-to-face interviews with show cards, (2) face-to-face interviews without show cards, and (3) telephone interviews with the same questionnaire as in group 2. After careful analysis, the findings suggest that differences were mainly due to the presence of the interviewer, causing a greater social desirability bias, and that show cards did not appear to affect responses.

3.5.3.3 Adjustment: Disentangling Mode and Selection Effects

Embedding a small randomized mode comparison experiment in a mixed-mode design makes assessing and compensating for mode effects much simpler. Assume that we have a mixed-mode design in which for reasons of efficiency different respondents are self-selected to two different modes. In this design, we embed a small-scale mode experiment where a small number of respondents are randomly assigned to one of these two modes. In this situation, we then effectively have four separate groups in our design. Figure 3.1 illustrates the four groups and the comparisons that can be made.

As Figure 3.1 shows, by embedding a small randomized experiment in the sampling, we can distinguish between the mode effect and the selection effect, and we can examine to what degree propensity score adjustment can correct for the selection bias. More importantly, if the difference in mode is actually a selection effect, we can ignore the apparent mode effect and interpret differences between respondents across modes as reflecting real differences. If the mode effect exists and is not negligible, we can use propensity score adjustment and attempt to equate the model parameters within the same mode

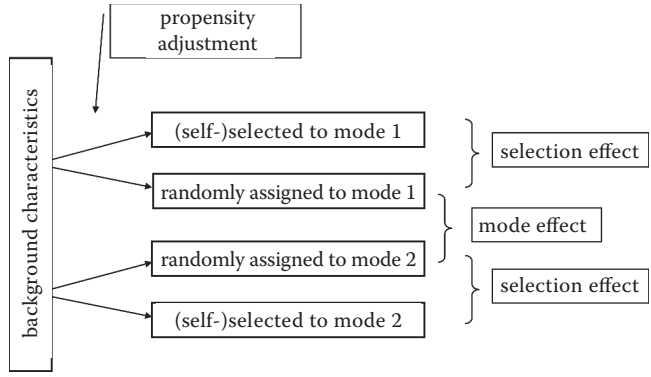

FIGURE 3.1
Combination of (self-)selection and random assignment to modes.

across the randomly assigned and the selected groups. The most principled way to accomplish this would be through structural equation modeling. In practice, however, after the researchers have investigated the various effects depicted in Figure 3.1 thoroughly, they may decide to use a simple dummy variable indicating mode in addition to the propensity adjustment.

In sum, when mixing Internet surveys with other data collection modes, much depends on the characteristics of the mix and the content of the questionnaire. Researchers should clearly evaluate the topic of the questions, the presence/absence of interviewers, and the communication channels (audio versus visual) used. Successful mixed-mode design requires an understanding of how mode differences may affect the answers given. For instance, when sensitive questions are used, the presence of an interviewer may cause social desirability bias associated with mode when CATI or CAPI is mixed with the more anonymous Internet survey. In this case, a mix of paper mail and Internet will be far better. When complex questionnaires are used, the aid of an intelligent system—be it a well-implemented Internet survey or an interviewer in CATI or CAPI—is necessary, and mixes with paper mail surveys may very well introduce respondent error in important routings. Also, the social conventions and customs associated with the mode are important for mixes involving Internet surveys, in particular when special groups (e.g., the elderly) are investigated or when cross-cultural and international surveys are being conducted. Finally, after having chosen the optimal Internet mix, designers must make sure that equivalent questionnaires

are implemented, as question-format effects have to be avoided. But even after careful designing, it is still possible that differences between modes will remain. To cope with these, it is useful to collect additional data on possible mode effects, or embed a small-mode experiment. These data are first used to investigate potential mode effects, and may be used later in the analysis phase to correct by statistical means for any mode differences.

REFERENCES

American Psychological Association. (1986). *Guidelines for computer-based tests and interpretations*. Washington, DC: American Psychological Association (APA).

Bergstrom, B. (1992). *Ability measure equivalence of computer adaptive and paper and pencil tests: A research synthesis.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco (as cited by Wang et al., 2007).

Bethlehem, J. G. (2008). Representativity of Web surveys—An illusion? In I. Stoop & M. Wittenberg (Eds.), *Access panels and online research: Panacea or pitfall?* (pp. 19–44). Amsterdam: Aksant/DANS. Retrieved from http://www.jelkebethlehem.nl/surveys/papers/bethlehem01.pdf

Biemer, P. P. & Lyberg, L. E. (2003). *Introduction to survey quality*. New York: John Wiley & Sons.

Blyth, B. (2008a). Mixed-mode: The only "fitness" regime? *International Journal of Market Research*, *50*(2), 241–266.

Blyth, B. (2008b). *The implications of variation in national data collection mode access and rate of access change: A European overview.* Paper presented at the 3MC conference, Berlin.

Bronner, F. & Kuijlen, T. (2007). The live or digital interviewer: A comparison between CASI, CAPI, and CATI with respect to differences in response behaviour. *International Journal of Market Research, 49*(2), 167–190.

Christian, L. M., Dillman, D. A., & Smyth, J. D. (2005). *Instructing Web and telephone respondents to report date answers in format desired by the surveyor*. Social and Economic Sciences Research Center Technical Report 05-067. Retrieved November 4, 2009, from http://survey.sesrc.wsu.edu/dillman/papers.htm

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. M. Lepkowski, C. Tucker, J. M. Brick et al. (Eds.), *Advances in telephone survey methodology* (pp. 250–275). New York: John Wiley & Sons.

Cole, M. S., Bedeian, A. G., & Feild, H. S. (2006). The measurement equivalence of Web-based and paper-and-pencil measures of transformational leadership: A multinational test. *Organizational Research Methods, 9*(3), 339–368.

Cook, C., Heath, F., & Thompson, R. L. (2000). A meta-analysis of response rates in Web- or Internet-based surveys. *Educational and Psychological Measurement, 60*(6), 821–836.

Couper, M. P. (2000). Web surveys: A review of issues and approaches. *Public Opinion Quarterly, 64*, 464–494.

Couper, M. P. (2008). *Designing effective Web surveys*. New York: Cambridge University Press.

Couper, M. P. & De Leeuw, E. D. (2003). Nonresponse in cross-cultural and cross-national surveys. In J. A. Harkness, A. J. R. Van de Vijver, & P. Mohler (Eds.), *Cross-cultural survey methods* (pp. 157–177). New York: John Wiley & Sons.

Couper, M. P., Kapteyn, A., Schonlau, M., & Winter, J. (2007). Noncoverage and nonresponse in an Internet survey. *Social Science Research, 36*(1), 131–148.

De Leeuw, E. D. (1992). *Data quality in mail, telephone, and face-to-face surveys*. Amsterdam: TT-Publikaties. Retrieved from http://www.xs4all.nl/~edithl

De Leeuw, E. D. (2005). To mix or not to mix data collection modes in surveys. *Journal of Official Statistics, 21*, 233–255. Retrieved from http://www.jos.nu

De Leeuw, E. D. (2008). Choosing the method of data collection. In E. D. De Leeuw, J. J. Hox, & D. A. Dillman (Eds.), *International handbook of survey methodology* (pp. 113–135). European Association of Methodology Series. New York: Lawrence Erlbaum Associates.

De Leeuw, E. D., Hox, J. J., & Dillman, D. A. (2008a). The cornerstones of survey research. In E. D. De Leeuw, J. J. Hox, & D. A. Dillman (Eds.), *International handbook of survey methodology* (pp. 1–17). European Association of Methodology Series. New York: Lawrence Erlbaum Associates.

De Leeuw, E. D., Hox, J. J., & Dillman, D. A. (2008b). Mixed-mode surveys: When and why. In E. D. De Leeuw, J. J. Hox, & D. A. Dillman (Eds.), *International handbook of survey methodology* (pp. 299–316). European Association of Methodology Series. New York: Lawrence Erlbaum Associates.

De Leeuw, E. D., Hox, J. J., & Kef, S. (2003). Computer-assisted self-interviewing tailored for special populations and topics. *Field Methods, 15*, 223–251.

Dillman, D. A. (2007). *Mail and Internet surveys: The tailored design method* (2nd ed.). New York: John Wiley & Sons.

Dillman, D. A. (2008). The logic and psychology of constructing questionnaires. In E. D. De Leeuw, J. J. Hox, & D. A. Dillman (Eds.), *International handbook of survey methodology* (pp. 161–175). European Association of Methodology Series. New York: Lawrence Erlbaum Associates.

Dillman, D. A., Brown, T. L., Carlson, J., Carpenter, E. H., Lorenz, F. O., Mason, R. et al. (1995). Effects of category order on answers to mail and telephone surveys. *Rural Sociology, 60*, 674–687.

Dillman, D. A. & Christian, L. M. (2005). Survey mode as a source of instability across surveys. *Field Methods, 17*, 30–52.

Dillman, D. A., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J. et al. (forthcoming). *Response rate and measurement differences in mixed mode surveys using mail, telephone, interactive voice response, and the Internet.* Retrieved from http://survey.sesrc.wsu.edu/dillman/papers.htm (earlier version).

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2009). *Internet, mail, and mixed-mode surveys: The tailored design method* (3rd ed.). New York: John Wiley & Sons.

Dillman, D. A., Smyth, J. D., Christian, L. M., & O'Neill, A. (2008). *Will a mixed-mode (mail/Internet) procedure work for random household surveys of the general public?* Paper presented at the annual conference of the American Association for Public Opinion Research (AAPOR), New Orleans, Louisiana.

Epstein, J. & Klinkenberg, W. D. (2001). From Eliza to Internet: A brief history of computerized assessment. *Computers in Human Behavior, 17*, 295–314.

Ferrando, P. J. & Lorenzo-Seva, U. (2005). IRT-related factor analytic procedures for testing the equivalence of paper-and-pencil and Internet-administered questionnaires. *Psychological Methods, 10*(2), 193–205.

Fricker, S., Galesic, M., Tourangeau, R., & Yan, T. (2005). An experimental comparison of Web and telephone surveys. *Public Opinion Quarterly, 69*(3), 370–392.

Groves, R. M. (1989). *Survey errors and survey costs*. New York: John Wiley & Sons.

Gwaltney, C. J., Shields, A. L., & Shiffman, S. (2008). Equivalence of electronic and paper-and-pencil administration of patient-reported outcome measures: A meta-analytic review. *Value in Health, 11*(2), 322–333.

Hacking, I. (1990). *The taming of chance*. Cambridge, UK: Cambridge University Press.

Heerwegh, D., Billiet, J., & Loosveldt, G. (2005). Opinies op bestelling? Een experimenteel onderzoek naar het effect van vraagverwoording en sociale wenselijkheid op de proportie voor-en tegenstanders van gemeentelijk migrantenstemrecht [Opinions on demand? An experimental investigation of the effect of question wording and social desirability on the proportion of proponents and opponents of municipal suffrage for immigrants]. *Tijdschrift voor Sociologie*, *26*(3), 189–208.

Heerwegh, D. & Loosveldt, G. (2008). Face-to-face versus Web surveying in a high-Internet-coverage population: Differences in response quality. *Public Opinion Quarterly, 72*(5), 836–846.

Heiser, W. (1996). *De probabilisering van het wereldbeeld [Chance and our view of the world].* Invited lecture, University of Groningen, symposium in honor of Professor Ivo Molenaar.

Jäckle, A., Roberts, C., & Lynn, P. (2007). *Assessing the effect of data collection mode on measurement.* Paper presented at the ISI conference, Lisbon, Portugal. See also ISER working paper 2006-41. Retrieved June 27, 2009, from http://www.iser.essex.ac.uk/publications/working-papers/iser/2006-41.pdf

Kim, J.-P. (1999). *Meta-analysis of equivalence of computerized and P&P tests on ability measures.* Paper presented at the annual meeting of the Midwestern Educational Research Association, Chicago (retrieved from ERIC).

Krauter, F., Presser, S., & Tourangeau, R. (2008). Social desirability bias in CATI, IVR, and Web surveys: The effect of mode and question sensitivity. *Public Opinion Quarterly, 72*(5), 847–865.

Krug, S. (2006). *Don't make me think: A common sense approach to Web usability: How we really use the Web.* Retrieved April 2009 from http://www.sensible.com/chapter.html

Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel Web surveys. *Journal of Official Statistics, 22*(2), 329–249. Retrieved from http://www.jos.nu

Link, M. W. & Mokdad, A. H. (2005). Effects of survey mode on self-reports of adult alcohol consumption: A comparison of mail, Web and telephone approaches. *Journal of Studies on Alcohol*, *66*, 239–245.

Lozar Manfreda, K., 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.

Lundstrom, S. & Sarndal, C.-E. (2002). *Estimation in the presence of nonresponse and frame imperfections.* Statistics Sweden.

Martin, E., Childs, J. H., DeMaio, T., Hill, J., Reiser, C., Gerber, E. et al. (2007). *Guidelines for designing questionnaires for administration in different modes.* Washington, DC: U.S. Bureau of the Census. Retrieved July 2009 from http://www.census.gov/srd/mode-guidelines.pdf

Mead, A. D. & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*(3), 449–458.

Meade, A. W., Michels, L. C., & Lautenschlager, G. J. (2007). Are Internet and paper-and-pencil personality tests truly comparable? An experimental design measurement invariance study. *Organizational Research Methods, 10*(2), 322–345.

Oosterveld, P. & Willems, P. (2003). Two modalities, one answer? Combining Internet and CATI surveys effectively in market research. In D. S. Fellows (Ed.), *Technovate* (pp. 141–150). Amsterdam: ESOMAR.

Preckel, F. & Thieman, H. (2001). *Testing intellectual giftedness on the Web: Development of a new figural matrices test-online versus paper-and-pencil versions.* Paper presented at the General Online Research conference (GOR '01), Göttingen, Germany. Retrieved April 2009 from www.gor.de

Richman, W. L., Kiesler, S., Weisband, S., & Drasgow, F. (1999). A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews. *Journal of Applied Psychology, 84*(5), 754–775.

Roberts, C. (2007). *Mixing modes of data collection in surveys: A methodological review.* ESRC/NCRM Methods Review Paper 008. Retrieved July 2009 from http://eprints.ncrm.ac.uk/418/1/MethodsReviewPaperNCRM-008.pdf

Rookey, B. D., Hanway, S., & Dillman, D. A. (2008). Does a probability-based household panel benefit from assignment to postal response as an alternative to Internet-only? *Public Opinion Quarterly, 72*(5), 962–984.

Schuman, H. & Presser, S. (1981). *Questions and answers in attitude surveys.* New York: Academic Press.

Shih, T.-H. & Fan, X. (2008). Comparing response rates from Web and mail surveys: A meta-analysis. *Field Methods*, *20*, 249–271.

Sikkel, D., Hox, J. J., & De Leeuw, E. D (2009). Using auxiliary data for adjustment in longitudinal research. In P. Lynn (Ed.). *Methodology of longitudinal surveys* (pp. 141–155). New York: John Wiley & Sons.

Smyth, J. D., Christian, L. M., & Dillman, D. A. (2008). Does yes or no on the telephone mean the same as check-all-that-apply on the Web? *Public Opinion Quarterly, 72*(1), 103–113.

Smyth, J. D., Dillman, D. A., Christian, L. M., & Stern, M. J. (2006). Comparing check-all and forced-choice question formats in Web surveys. *Public Opinion Quarterly, 70*(1), 66–77.

Sudman, S. & Bradburn, N. M. (1974). *Response effects in surveys: A review and synthesis*. Chicago: Aldine.

Toepoel, V. (2008). *A closer look at Web questionnaire design*. CentER for Economic Research Dissertation Series, no. 220, Tilburg University, the Netherlands.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response.* Cambridge, UK: 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*, 275–304.

Turner, C. F., Forsyth, B. H., O'Reilly, J. M., Cooley, P. C., Smith, T. K., Rogers, S. M. et al. (1998). Automated self-interviewing and the survey measurement of sensitive behaviors. In M. P. Couper, R. P. Baker, J. Bethlehem et al. (Eds.), *Computer-assisted survey information collection* (pp. 455–473). New York: John Wiley & Sons.

Van Buuren, S., Eyres, S., Tennant, A., & Hopman-Rock, M. (2005). Improving comparability of existing data by response conversion. *Journal of Official Statistics*, *21*, 53–72.

Vandenburg, R. & 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.

Van Ewijk, R. (2004). Onderzoek via telefoon en Internet: De verschillen [Surveys by means of telephone and Internet: The differences]. *Clou, 14*, 38–40.

Van Meurs, A., Van Ossenbruggen, R., & Nekkers, L. (2009). Rotte appels? Controle op kwaliteit van antwoordgedrag in het Intomart GfK Online Panel [Do rotten apples spoil the whole barrel? Checking the quality of responses in the Intomart GfK Online Panel]. In A. E. Bronner et al. (Eds)., *Ontwikkelingen in het Marktonderzoek, Jaarboek Marktonderzoek Associatie, 34*, 61–81.

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2007). A meta-analysis of testing mode effects in grade K-12 mathematics tests. *Educational and Psychological Measurement, 67*(2), 219–238.

Wang, S., Jiao, H. Young, M. J., Brooks, T., & Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing in K-12 reading assessments: A meta-analysis of testing mode effects. *Educational and Psychological Measurement, 68*(1), 5–24.

Willems, P., Van Ossenbruggen, R., & Vonk, T. (2006). The effect of panel recruitment and management on research results: A study across 19 online panels. *Proceedings of the ESOMAR World Research Conference, Panel Research 2006, 317*, 79–99. Amsterdam: ESOMAR.

Wine, J. S., Cominole, M. B., Heuer, R. E., & Riccobono, J. A. (2006). *Challenges of designing and implementing multimode instruments.* Paper presented at the Second International Conference on Telephone Survey Methodology, Miami, Florida. Retrieved March 2009 from http://www.rti.org/pubs/TSM2006_Wine_paper.pdf