7

Survey Data Collection

Modes of survey data collection just refer to the different forms of communication used to contact and solicit respondents, ask questions and record the answers. The mode affects a respondent's willingness to answer a survey and his or her ability and motivation to understand and answer answers. *Interviewer effects* on survey response arise from subtle differences in the way that interviewers and record responses as well as respondents' perceptions of interviewers.

The chapter begins with a description of the four principal modes of data collection – personal and telephone interviews and 'paper and pencil' and Internet self-administered surveys – and then looks at the effect of survey mode on response rates and data quality and the impact of interviewers. The second part of the chapter examines the factors affecting the survey response rate and the impact of response rates on estimates of population characteristics.

To a greater degree than sample and questionnaire design, survey data collection is shaped by its local context, and its success depends critically on respondents' understanding of and sympathy with the survey enterprise. For example, telephone surveys are a cost-effective alternative to face-to-face household surveys, but only if: nearly everyone can be reached by telephone; it is possible to draw a sample of numbers; and most people will answer the survey calls and agree to be interviewed.

At this time, any discussion of data collection takes place in the shadow of a widely perceived crisis in the ability to gather data adequately representing a 'general' population. Recently, Groves (2011) characterized the present period of survey research, beginning in 1990, largely in terms of declining response rates coupled with and in many respects caused by advances in technology. While data collection is hugely varied – a function of the survey sponsor, research topic, data collection agency and the length and demands of the survey – declining response rates are a general, formidable and worsening problem. Rationalizing the decline in response rates, there has been a revision of the traditional view that a high response rate is a necessary criterion of a good survey. Instead the emphasis has

shifted to measuring the impact of non-response on estimates of population characteristics. These issues are also taken up in the next chapter on the future of surveys.

Modes of Data Collection

Traditionally, a mode of data collection combined a method of contacting respondents with a method of asking questions and recording responses. For example, respondents would be contacted and interviewed by telephone or they would be contacted by email and respond on the Internet. But other combinations are possible and have become increasingly common. Respondents may be contacted by telephone and asked to complete an Internet questionnaire or face-to-face interview, or they may be contacted in more than one way and be able to complete a survey in more than one mode.

Face-to-Face Surveys

Because this was the only way to select a probability sample and administer questionnaires with an adequate response rate, surveys began with face-to-face interviews. They continue to be used when there is no other way to select a representative sample and when other methods do not result in an acceptable response rate, including mandated government surveys, surveys over about 20 minutes in length and surveys in locales or regions where a substantial part of the population cannot be surveyed by telephone. Because of their much higher cost, face-to-face surveys are used when no other mode will provide the required data.

As laptops became more capable and less expensive, face-to-face interviewing moved from paper and pencil questionnaires to computers, using portable versions of the software first developed for centralized telephone surveys. Computer-assisted personal interviewing (CAPI) both eliminates the cost, delay and errors introduced by a separate data entry step and prevents errors in navigation between questions. Computerization also allows complex sequences of questions tailored to the characteristics of respondents, facilitates experiments in question wording and order, and can be used to provide the interviewer with additional information as the survey proceeds, such as the answers to questions commonly raised by respondents and advice on dealing with unusual responses.

The drawback of face-to-face surveys is their cost. Locating and interviewing a respondent can require several trips, because no one is home or the selected respondent is not present or will not do an interview at the time. Controlling costs requires limiting the number of calls to each address, because the likelihood of obtaining an interview declines with the number of unsuccessful previous calls. Almost always cost considerations require the selection of clusters of addresses, typically 10 to 15, in small geographical areas, and it is also common to exclude remote or sparsely populated areas altogether. This clustering decreases

the precision of estimates because neighbours tend to resemble one another in socio-demographic characteristics and in other ways.

Face-to-face surveys can easily incorporate alternative modes of response. Small printed cards called 'show cards' may be used to present the response options for a complicated question or respondents may be asked to sort cards or carry out other tasks with materials provided by the interviewer. For sensitive questions, respondents can be given a printed form to complete in private and return to the interviewer in a sealed envelope, or respondents' may be asked to use the interviewer's computer themselves, for privacy or to answer questions with a visual component. The latter is known as computer-assisted self-interview, or CASI, and A-CASI is the same, except the respondent listens to questions with headphones (the A is for 'audio'). Adding another mode of data collection to a face-to-face survey does not make it a 'mixed-mode' survey, because each question is answered using one mode.

In longitudinal surveys a common strategy is to use an initial face-to-face interview to obtain a higher response rate and conduct a longer initial interview and, perhaps most important, develop rapport with a respondent whose long-term commitment to the survey is critical. In later waves, data collection may be switched to a telephone or perhaps an Internet survey, but continue with face-to-face interviews for respondents who cannot or will not use another mode. An initial face-to-face interview is also a good time to ask a respondent for permission to access health, taxation and other administrative records and to obtain the names and coordinates of contacts who could locate the respondent if she moves.

Telephone Surveys

Telephone and face-to-face surveys are similar in that an interviewer is able to select a respondent within a household and ask questions and record the answers, but telephone sample coverage is generally not as good. Almost everyone has an address, but even where telephones are almost universal, some people have no telephone, never answer calls, or reject any survey request over the telephone. Compared to a face-to-face survey, the anonymity of the telephone makes it difficult for an interviewer to tailor the invitation to the characteristics of a potential respondent and also makes it easier for the respondent to refuse.

Typically, each telephone interview is one-third to one-fifth the cost of a comparable face-to-face interview. Not only are travel costs eliminated, but the centralization of telephone surveys provides efficiencies in the organization and supervision of interviewing. Because telephone numbers are generally not clustered geographically,[1] a telephone sample has lower sampling error than the

[1]Even with traditional switching systems, the addresses reached by telephone numbers that differ by only one digit are not sufficiently close to result in significant clustering, and the adoption of electronic telephone switching removes any need to associate telephone numbers with geographical areas.

same-size face-to-face survey. For variables with a significant degree of neighbourhood variation a telephone survey might have the precision of a face-to-face survey with 20 to 40 per cent more observations.

Self-administered Surveys

To conduct a self-administered survey of the general population requires a sampling frame listing individuals by name. It cannot be assumed that instructions about how to select a household member at random, which are enclosed with a survey with only a household address, will be followed correctly by whoever opens the envelope. This is no problem if the questions pertain to the entire household and can be answered by any competent household member, but then it is difficult to obtain an adequate response rate if a mailed survey has only an address. Self-administered surveys are often used to survey organizations and businesses, by postal mail or over the Internet, and in a few countries the availability of a population register makes it possible to conduct mail surveys with a probability sample of the population.

The cost of a mailed survey is somewhat lower than a comparable telephone survey, depending on the efforts made to contact selected respondents. After the start-up cost, mail surveys have no economies of scale because the survey costs are mainly printing, postage and the computer entry of completed questionnaires. Internet surveys involve a substantial start-up cost, especially for visual design, programming and testing with the more advanced proprietary software used by major survey organizations, but except for any incentives their cost is essentially unrelated to the sample size . For researchers willing to format their own Internet questionnaires and live with the limitations of the software of commercial online providers, Internet surveys are very low in cost.

Because self-administered surveys allow respondents to answer at their own pace, they can include more complex questions than surveys employing interviewers. Especially compared to telephone interviews, respondents' answers are less affected by their ability to remember the details of a complicated question or keep in mind a number of possible answers. Self-administered questionnaires can make use of the visual design of questions, including illustrations, but the layout can also introduce bias. Compared to surveys conducted by an interviewer, order effects are diminished when the respondent is able to see a number of questions on the same printed page or screen and also because the respondent can double back and change the answers to earlier questions. In self-administered questionnaires, the answers to adjacent questions are likely to be more consistent and less prone to order effects.

While a self-administered questionnaire allows the respondent the time to answer questions carefully, without an interviewer there is no social pressure to discourage carelessness and encourage the respondent to answer every question. The distracting environment of multi-tasking computers could also lead to less accurate response.

Interactive voice response (IVR) surveys are a hybrid between telephone and self-administered surveys. Contact is made by telephone, with automated dialling or the respondent calling in, and respondents answer by using the telephone keypad or by voice. There are no interviewer effects, since every respondent hears the same pre-recorded voice and the responses are recorded automatically. While IVR response rates tend to be low and the survey must be very short, the cost is also very low. IVR is increasingly used for quick polls during an election campaign.

Multi-mode Surveys

Since the mid-2000s survey researchers have increasingly used multi-mode surveys, in order to lower costs, but more to address declining response rates in household telephone surveys. Many combinations of the modes of recruitment and data collection are possible (de Leeuw, 2005: 237ff.). For example, with a sample of addresses, respondents may be contacted in person, by telephone if a number can be linked to the address, by postal mail and potentially by email. The survey can be administered by an interviewer, in person or by telephone, or the respondent can complete a mailed or Internet questionnaire.

Savings are possible if lower cost, self-administered questionnaires can substitute for interviews or if telephone interviews take the place of face-to-face interviews. The idea is to combine the better coverage of address samples with the lower cost of self-administered questionnaires, reserving interviews for respondents who will not otherwise respond. These cost differences may be sufficient to pay for moderate incentives to respondents who use a lower cost mode. For an interesting discussion of the European Social Survey (ESS) changing from face-to-face interviews to mixed-mode form, see Martin (2011: 6ff.).

Net of the characteristics of respondents, the mode of data collection itself can give rise to differences in response. Because the choice of mode is likely related to respondent characteristics, this could result in bias in the estimates of population characteristics and between-group differences. This is the subject of much current research (for example, see de Leeuw et al., 2010 and Vannieuwenhuyze et al., 2010). The issue is complicated because characteristics of the respondent, *not* all of which are measured in a survey, are likely to affect the *mode* chosen by the respondent *and his or her survey responses*, so it may not be possible to fully remove the mode-related bias.

Except where there are known and large mode effects, for example surveys on alcohol consumption, the consensus is that the benefits of lower cost and higher response outweigh the risks of undiscovered mode-related bias. Also, single-mode surveys are not free of mode bias – it is just that any such bias cannot be measured. Mode effects in 'ordinary' questions are typically quite small, and no larger than other unavoidable errors that, for example, arise from the choice of nearly synonymous words in writing a question, the order of response options and the respondents' occasional and unsystematic misunderstandings.

Survey Mode and Response Rates

The two most comprehensive comparisons find that face-to-face surveys have higher response rates[2] than telephone surveys, by a small margin. De Leeuw (1992: 26) reports a 75 per cent mean response rate for face-to-face surveys, versus 71 per cent for telephone surveys and Goyder (1987: 56) obtains similar results, after accounting for the number of contacts with a respondent, the use of incentives and the survey topic and sponsor. Especially in light of the very large cost difference, this evidence lent support to the shift from face-to-face to telephone surveys from about the mid-1970s.

These findings, however, long predate the decline in landline coverage in a number of countries, the difficulties in sampling and completing interviews with mobile phones and the widespread adoption of 'call screening', not to mention the overall decline in response rates. For the USA, Holbrook et al. report that 'a diverse group of the nation's most expensive, long-term, large-scale, federally funded survey studies of national samples involving long questionnaires have retained the face-to-face method while most other survey research moved to the telephone' (2003: 81).

In recent meta-analyses, Lozar Manfreda et al. (2008) and Shih and Fan (2008) found average response rates of about 45 per cent for mail surveys and 35 per cent for Internet surveys. While the 10 per cent average difference is quite large, this does not mean that a mail survey will invariably result in a higher response rate, because the standard deviation in the response rates of the surveys using each mode is about 20 per cent. In choosing the survey mode, cost is likely to be a big factor. For a small sample, the effort required to program an Internet survey might be similar to the resources needed to design, print and mail a survey and enter the returned forms. For a larger sample, however, the printing, postage and data entry costs of a mailed survey make the Internet survey much less expensive.

Because survey response can only be compared for samples of the same type, there are not many comparisons between self-administered and interview

[2]In principle, the response rate is just the number of completed questionnaires divided by the number of selected respondents. But there is some uncertainty in both the numerator and denominator. For the numerator, we must decide what constitutes a completed questionnaire. Some respondents quit before the end or complete the questionnaire but leave many questions unanswered. For the denominator, we must determine whether selected respondents who we never contacted are *eligible* respondents. A conservative estimate of the denominator is just the average eligibility for respondents *who were successfully contacted*. Alternatively, an estimate can be obtained by regressing the probability of eligibility on auxiliary variables, such as location, that are available for all selected respondents and on measures of the difficulty of contacting respondents, such as the required number of calls. To encourage uniform and fair assessments, as well as create a level playing field for researchers and survey organizations that would prefer to report higher response rates, there is wide acceptance of the expert guidelines of the American Association for Public Opinion Research (2011). It defines six alternative response rates, based on whether partially completed surveys are counted and the assumptions made about the eligibility of selections who were not contacted.

surveys. Fricker et al. (2005) and Chang and Krosnick (2009) obtain a higher response rate for a telephone survey than an Internet survey, but their comparisons are based on a sample initially contacted by telephone, so the telephone respondents are immediately asked to continue with a telephone survey while the Internet group is asked to log on at some later time. Contacted on the telephone, more respondents will proceed with an interview immediately, or maybe arrange a time to call back, than will agree to and then fulfil their commitment to answer an Internet survey.

It follows that offering respondents a choice of mode *when they are first contacted* does not improve response rates, again because of the failure to follow through on promises to complete a self-administered survey. If the alternative modes are offered *in sequence*, however, some respondents who refuse to answer an interview will complete a self-administered survey (de Leeuw, 2005: 240), which produces some improvement in the response.

There is considerable international variation in response rates. For 21 countries, Stoop et al. (2010: 93) report that response rates for the 2006 ESS, a biennial face-to-face survey, varied between 46 and 73 per cent. Within nations, however, there are also differences between modes. While the ESS response rate for the Netherlands is 59.8 per cent, a bit below average, the elaborate *Internet* LISS panel study, based on a probability sample drawn from the Netherlands national register, has a response rate around 80 per cent.[3] This reversal of the usually higher response rate of face-to-face surveys must in part be due to the LISS panel's use of incentives and its providing a free computer with Internet access to households without one. The average cost of ESS face-to-face interviews is €120, which would provide for a healthy incentive if it was possible to switch to a self-administered survey.

Survey Mode and Data Quality

As the total survey error perspective emphasizes, like each of the other steps in the survey process, the mode of data collection contributes to measurement error. The two common forms of mode-related bias, *acquiescence* and *social desirability*, were discussed in Chapter 2, in relation to question design.

[3]See the ESS at http://www.europeansocialsurvey.org and the LISS project at http://www.lissdata.nl. Another large international survey project, the International Social Survey Program (ISSP), allows the researchers in each country to decide the mode of data collection and does not provide response rate comparisons. Surveys for the 'EU statistics on income and living conditions' (EU-SILC) are conducted by the different EU nations' central statistical agencies, also using a variety of modes, including multiple modes in some nations, to study income, poverty, social exclusion and living conditions. Response rates in 2009 varied from 52 to 96 per cent in the 32 participant countries (Eurostat, 2011: 27).

More concern has been directed to the effect of survey mode on random error, non-response and response patterns indicative of satisficing and inattention. A number of differences have been observed, for example:

- because they cannot see the interviewer or read questions themselves, telephone survey respondents have the most difficulty understanding and answering complex questions;
- self-administered surveys allow respondents to set their own pace and also it is easier to reread a difficult question than to ask an interviewer to repeat it;
- compared to the telephone, face-to-face interviewers communicate more effectively and are better able to sense and try to clarify respondent confusion;
- face-to-face interviews and self-administered surveys permit the use of visual material on a show card, printed page or monitor;
- interviews generally, and especially telephone interviews, put more burden on the respondent's ability to remember and consider the details in a question and to consider all the response categories; and
- only interview surveys allow respondents to ask for help, though there is evidence that they do not often do so. (Conrad and Schober, 2005: 221)

The environment and technology of telephone interviews and Internet surveys seem to encourage less thoughtful responses than face-to-face interviews or paper and pencil surveys. Answering a survey on the telephone at home or a mobile phone, the respondent is in a potentially distracting environment where it is easy to surreptitiously multi-task. Modern computer environments, where the Internet survey is just in one 'window', might also encourage quicker, less thoughtful responses.

The possibility that the survey mode affects a respondent's motivation and the quality of her or his responses is captured by the idea of survey *satisficing*, put forward by Jon Krosnick (1991; also see Krosnick and Alwin, 1987). First identified and given the name by Herbert Simon in 1957, satisficing describes decision making that aims for a good and acceptable outcome, rather than the optimal outcome.[4] Applying the concept to survey respondents, Krosnick distinguishes 'weak' and 'strong' satisficing, respectively, as:

> being less thorough in comprehension, retrieval, judgement, and response selection … [being] less thoughtful about a question's meaning, they may search their memories less thoroughly, they may integrate retrieved information more carelessly, and they may select a response choice more haphazardly;
>
> omitting the retrieval and judgement steps from the response process altogether. That is, respondents may interpret each question only superficially and select what they believe will *appear* to be a reasonable answer to each question. (1991: 215)

[4]Satisficing makes sense in many aspects of everyday life, and it plays a role in more important personal decisions when it is hard to judge outcomes exactly or where the cost of securing an optimal outcome is greater than the expected gain over a pretty good outcome. For example, job seekers may curtail their search for a better job when they find an acceptable one, because finding a better alternative takes time and after an initial substantial search is likely to result in only small gains.

Satisficing results in increased item non-response and in the proportion of neutral and middle answers, as well as patterns of *non-differentiation*, whereby a respondent gives the same answer to a series of consecutive questions, regardless of their content.

Ideally, mode effects would be measured by comparing survey responses to factual external information. Even in the rare circumstance when this is possible, however, validation studies are expensive and difficult, and often they involve special populations from which it is difficult to generalize. For attitudes and many other common survey measures, of course, there is no such external reference and so mode effects are measured with experiments were respondents are assigned to a survey mode at random. In principle, any survey with mixed-mode data collection provides a basis for comparing modes, but such comparisons are often compromised by incompletely measured differences between respondents using the different modes. Mode effects are especially difficult to measure when respondents are first asked to answer in the lowest cost, self-administered mode, and then switched to more costly modes if they do not respond.

Empirical Comparisons

Accompanying and rationalizing the growth of telephone surveys in countries with good landline coverage in the 1970s, a large body of research compared face-to-face and telephone surveys. At the time response rates were quite high for both modes, typically around 70 per cent. In the most extensive comparison, which was conducted by Groves and Kahn (1979), there was no evidence of systematic differences, with two exceptions:

- For questions that used seven *labelled* categories, telephone respondents were more likely than face-to-face respondents to choose the middle category and the extreme positive category, and less likely to choose the extreme negative category; using seven points labelled only at the ends and in the middle, however, there was virtually no difference between modes (p. 148).
- For open questions that allowed multiple responses, telephone respondents gave fewer answers and the telephone and face-to-face response distributions differed (p. 149).

Ye et al. (2011) find similar 'extremity bias' in telephone surveys, compared to both self-administered questionnaires and face-to-face interviews, but only for positive answers. What they call 'The MUM effect [that is, 'keeping mum'] applies to situations in which the message is undesirable for the receiver. This is quite distinct from the concept of social-desirability bias, which refers to respondents' reluctance to reveal embarrassing information about themselves' (p. 351).

Summarizing a number of studies, de Leeuw (1992: 28) finds that, to a small degree, telephone surveys exhibit more item non-response and they are more

likely to elicit similar responses to a series of consecutive questions (indicating satisficing) and fewer responses to open questions. She finds no difference in social desirability in surveys conducted in the 1980s, but slightly more social desirability in telephone surveys prior to 1980. Comparing face-to-face and telephone interviews in US surveys from 1976, 1982 and 2000, Holbrook et al. found that telephone survey respondents were more likely 'to satisfice (as evidenced by no-opinion responding, nondifferentiation, and acquiescence), to be less cooperative and engaged in the interview … [and] to present themselves in socially desirable ways' (2003: 79).

Also comparing face-to-face to telephone interviews in an experiment, Jäckle et al. (2010) found that telephone interviews elicited more socially desirable responses for 16 out of 28 variables tested, including moderate, statistically significant effects (around 5 per cent in magnitude) on reported interest in politics and political efficacy, religiosity and attendance at religious services, household income and large effects (around 15 per cent) on the total time spent watching television and watching television news, and attitudes towards immigration. They also showed that for typical ordinal measures of attitudes and behaviour, the estimated magnitude of the mode effect depends strongly on the statistical model employed (p. 7ff.).

The extent of mode effects is related to the question topic. In a large Canadian health survey, Béland and St-Pierre found no difference between face-to-face and telephone interviews for 'the vast majority of health indicators … such as tobacco use (all ages), chronic conditions, activity limitations, fruit and vegetable consumption' (2008: 6). There was evidence of social desirability bias, however, in telephone responses to questions about height, weight, physical activity, contact with physicians and unmet health care needs.

There is also evidence of international differences. Comparing face-to-face and CATI interviews, Martin finds that

> in the ESS CATI experiment … telephone respondents were more likely than face-to-face respondents to use extreme points on response scales, while face-to-face respondents were more likely to agree with the premise of questions (acquiescence effects) and to use scale mid-points. Importantly, the results were not uniform in all participating countries. (2011: 11)

Comparisons between interviews and self-administered surveys demonstrate one advantage and one disadvantage of each. First, there is extensive evidence that self-administered surveys, whether mail or Internet, result in more truthful answers to sensitive questions than face-to-face and telephone surveys (de Leeuw, 1992: 30; Aquilino, 1994; Tourangeau and Smith, 1996; Tourangeau and Yan, 2007: 863; Krauter et al., 2008). On the other hand, self-administered surveys by mail and especially Internet surveys show evidence of more satisficing in the form of item non-response, random error and non-discrimination (Fricker et al., 2005; Heerwegh and Loosveldt, 2008). Chang and Krosnick (2009: 641) report that a telephone survey 'manifested more random measurement error, more

survey satisficing, and more social desirability response bias' than their comparable Internet survey.

In terms of data quality, the clear winners are face-to-face interviews over telephone interviews, and paper and pencil self-administered surveys over Internet surveys. For most surveys, which employ a single mode of data collection, the evidence of mode effects can only be cautionary. Except for sensitive topics which demand self-administered surveys, the choice of survey mode is almost always determined by the ability to sample the population of interest, budgetary considerations and sometimes the length of the survey questionnaire.

In multi-mode surveys, mode effects may give rise to methodological artefacts because of the correlation between mode and respondent characteristics. This is also a concern for comparative surveys where modes vary across nations and for surveys employing multiple modes in succession in order to increase response rates or reduce costs.

The correlation between survey mode and respondent characteristics also affects the analysis of mixed-mode data. Unless respondents are assigned to modes randomly, which is rare unless the surveys are designed explicitly for methodological research, it is a good idea to include questions that predict the mode preferred by a respondent. If possible, a better strategy is to embed an experiment that randomly allocates a subsample of respondents to the different modes.

A special case is longitudinal surveys that switch from interviews in the first wave(s) to less expensive self-administered modes. Then it is often possible to draw on previous research and effectively account for the impact of changing modes. If the resources are available, the gold standard is to gather data with both methods in the wave when the transition takes place, with at least part of the sample assigned to modes randomly. A similar strategy is commonly employed when official surveys change the wording of critical questions.

Minimizing mode effects in multi-mode surveys can come into conflict with maximizing the quality of questions in each mode. For example, a technique known as unfolding can be used in interviews to increase the precision of response. The respondent is first asked whether he or she agrees or disagrees with a statement, and then a second question asks whether he or she (dis)agrees or *strongly* (dis)agrees. This makes up for the inability of the interview respondent to see all the alternatives easily on paper or a screen, but could account for at least a small difference between modes. Similarly, since the visual design of a question can affect response, it can also give rise to a mode effect if some observations are collected by interview.

There is no perfect solution to the problem of designing questionnaires for use in more than one mode, but Dillman (2007: 232ff.) makes a convincing case for *unimode construction*, which he defines as 'writing and presenting … questions to respondents in a way that assures receipt by respondents of a common mental stimulus, regardless of the survey mode'. The question is to what degree it is possible to present respondents with 'a common mental stimulus' in light of

the cognitive differences between modes, as well as differences in the setting. De Leeuw is more tentative:

> Hardly any theoretical or empirical knowledge is available on how to design optimal questionnaires for mixed-mode data collection (e.g., unimode and generalized mode design). Empirical research is needed to estimate what constitutes the same stimulus across different modes, and especially how new media and new graphical tools will influence this. (2005: 249–250)

Likewise, there is a need for the further development of analysis of multi-mode surveys. Overwhelmingly, the existing literature is concerned with the use of multi-mode surveys to increase, or least arrest the decline in, response rates, rather than with the design of multi-mode questionnaires or analysis of the resulting data.

Interviewer Effects

Interviewers do what letters, emails and invitations on websites cannot: they actively convince respondents to participate, often arranging for a return call or visit; they select respondents within a household; and they answer questions during an interview. There is also variation between interviewers in the ability to secure cooperation, read questions and record answers accurately, and in how they answer queries from respondents. Training and supervision designed to standardize interviewers' behaviour cannot completely eliminate the differences arising from interviewers' personalities, skills and experience and it cannot control the effects of *respondents'* perceptions of the voice and appearance of interviewers.

Since surveys are intended to capture variation in respondents' answers to identical questions, *any* effect of an interviewer on respondents' answers is a form of measurement error. Such error can be classified into two categories. First, in their role as intermediaries between a researcher's questions and respondents' answers that they record, interviewers contribute some *random* error, adding to errors arising from respondents' varying ability to understand and answer questions and the effort put into their answers. Interviewer random error arises from minor misreading of questions and unintended emphases and mistakes in recording responses; it is difficult to separate from the other error contributions, but it is not problematic. The second form of interviewer effect involves systematic differences between the groups of respondents allocated to each interviewer – in statistical terms these constitute clusters – that cannot be attributed to the attributes of the respondents in each cluster. 'Interviewer effects' refer to this second component only.

What should the interviewer do when a respondent clearly misunderstands a question or asks for clarification? If she steps in to correct the misunderstanding or answers a question, the result is a better answer. But another interviewer might respond differently and so the different styles of interviews can give rise to interviewer effects. This is why it is common for survey organizations to attempt to reduce interviewer variation by instructing the interviewer to do no more than

reread the question when a respondent asks for clarification, even if this does not help the respondent give a better answer. So there is potentially a trade-off between interviewer- and respondent-related error.

Measuring Interviewer Effects

Like clusters of respondents that arise from the sample design, interviewers potentially increase the similarity of responses of the respondents in the same cluster. This is measured by the intraclass correlation due to interviewers, symbolized ρ_{int}. If each interviewer conducts an average of m interviews, the error variance is increased by a factor of $\text{deff}_{int} = 1 + \rho_{int}(m - 1)$, where deff_{int} is called the design effect. So, if ρ_{int} is zero or if each interviewer does just one interview, there is no interviewer effect!

Say that $\rho_{int} = 0.02$ for some question and each interviewer conducts an average of 30 interviews. Then, from the formula above, deff_{int} is 1.58, meaning that the allocation of groups of respondents to interviewers results in a 58 per cent increase in the error variance and a 26 per cent increase (since $\sqrt{1.58} = 1.26$) in the confidence interval of a mean or proportion. This is substantial, since it means that making up for the loss in precision due to the interviewer effect would require a 58 per cent increase in the sample size. Clearly the aim of training and supervising interviewers should be to minimize the variation in interviewers' behaviour that increases the value of ρ_{int}. The magnitude of the interviewer effects is different for each variable in a survey, and is potentially a function of the survey population, mode, survey topic and style of interviewers.

Following Kish (1962; also see Hartley and Rao, 1978), the intraclass correlation is equal to the proportion of the total variance in a variable that is due to variation between interviewers. If respondents were randomly assigned to interviewers, estimating ρ_{int} would require only a one-way analysis of variance (ANOVA). In face-to-face surveys where each geographical cluster is assigned to a single interviewer, the geographical and interviewer effects cannot be separated. Separating them requires an *interpenetrated* design, whereby the addresses in each cluster are divided randomly between at least two interviewers. This raises travel costs because it decreases the likelihood that each interviewer's travel to a cluster will result in an interview, though it is more feasible in cities where the population density is higher. With an interpenetrated design, interviewer effects *net of geographical clustering* can be estimated with a two-way ANOVA or, more elegantly, with a mixed model.

For centralized telephone surveys it is quite easy to allocate respondents to interviewers at random, though it is necessary to suspend the common practice of assigning the most skilled interviewers to 'convert' respondents who refuse an initial request and deal with respondents who repeatedly delay but never refuse to complete an interview outright.

Because of the effort required to estimate the interviewer effects, the number of reported studies is a small and not-likely representative sample of all surveys. For the 12 pre-1985 studies located by O'Muircheartaigh and Campanelli (1998: 68), the average of the value of ρ_{int} for face-to-face interviewers is about 0.01. So, if

each interviewer averaged 40 interviews, $deff_{int}$ was about 1.4. They also contribute a nice analysis of the 1992 wave of the British Household Panel Survey, estimating ρ_{int} for no less then 802 variables. The median ρ_{int} is about 0.01, but there is substantial variation; ρ_{int} is significantly different from zero for 28 per cent of the attitude questions and 26 per cent of the factual questions.[5] The value of ρ_{int} was unrelated to whether the question was from the 'cover sheet' (completed by the interviewer and recording household composition and similar measures), or whether questions related to the individual respondent or the respondent's household. Moreover, 17 per cent of the items that respondents answered privately on a printed form also had ρ_{int} values significantly greater than zero. These results are consistent with results reported by Schnell and Kreuter (2005).[6]

Interviewer effects are smaller in telephone than face-to-face surveys, although the two largest comparisons, both American, are quite dated. In analysing 1980 election polls, Tucker (1983) found an average value for ρ_{int} of 0.04. Groves and Magilavy (1986) report that the average value of ρ_{int} was 0.009 for nine surveys conducted between 1978 and 1982, with slightly higher values of ρ_{int} for factual items, 0.0098, than for attitudinal items, 0.0085. More recently, Lipps (2007) found negligible interviewer effects in a Swiss telephone survey.

A fine summary of research by Davis et al. (2010) provides unequivocal evidence that visible and audible characteristics for the interviewer, most notably gender and racialization, affect the answers to questions *related to those statuses*. For example, both women and men are more likely to endorse women's rights if the interviewer is female. The argument can be made that the interviewer and respondent should be matched on the assumption, for example, that women respondents are most likely to give unbiased responses to women interviewers and the same is true for men. In reviewing the empirical evidence, however, Schaeffer et al. (2010: 443) find no support for this conjecture. Matching interviewers and respondents is difficult in household surveys because the characteristics of the respondent are not known before contact with the household. An easier strategy is to control for characteristics of the interviewer in data analysis, provided they are sufficiently diverse (a problem if the interviewers are predominantly women).

Minimizing and Accounting for Interviewer Effects

If interviewer effects potentially result in bias, how can they be minimized? The conventional strategy is greater 'standardization', insisting that questions be read exactly as written, limiting responses to respondent queries to rereading the question, and asking respondents to answer in terms of what 'the question means to

[5]But, as usual, a non-significant result is not evidence that the value is zero.

[6]Analysing a recent German survey, Schnell and Kreuter obtain a mean design effect of 1.39 for geographical clusters and find that 77 per cent of the between clusters is due to interviewers and 23 per cent to the sampling point (pp. 400–401). The magnitude of the effects is slightly higher for sensitive items, non-factual items and open questions, but is unrelated to the difficulty of items (2005: 402).

you'. In a recent review, however, Schaeffer et al. (2010) conclude that standardization is unlikely to significantly decrease interviewer effects. This makes sense because the observed design effects are computed from surveys conducted by large and professional organizations whose interviewers were already highly trained.

A number of aspects of the working conditions of interviewers militate against the likelihood that standardization will decrease interviewer effects. Often interviewers work part-time and are poorly paid. The resulting high turnover means that investment in more intensive training largely evaporates. Second, a heavy emphasis on standardization makes the job less pleasant, which may lower the quality of work, decrease productivity and increase turnover. Third, for most survey organizations an interviewer's most important skill is convincing respondents to be interviewed, which requires an outgoing but not overbearing personality that might not fit with a highly regimented approach to conducting interviews.

The literature on survey research displays disturbingly little interest in the quality of the workplace and its effect on the quality of survey data. Stoop et al. comment:

> Considering the importance of the role of the interviewer, it is key that their payment reflects the efforts required. Levels of interviewer pay and the pay structure are highly likely to affect interviewers' willingness to try and enhance their response rates … Payment systems, assignment sizes, training programmes and other practical issues are specific to each fieldwork organization. Differences between countries are to be expected. No empirical studies are available on the effect of such differences. (2010: 23)

Excessive standardization of interviews may actually lower the quality of responses, because the rigidity alienates respondents and the interviewers do not help the respondents give better answers. Thus decreasing the measurement error due to interviewer effects may increase the random measurement error, because it restricts the ability of interviewers to help respondents give better answers.

A number of methodologists (Beatty, 1995; Schober and Conrad, 1997; Conrad and Schober, 2000; 2005; Maynard and Schaeffer, 2002) advocate greater efforts to clarify the meaning of questions, including improving the explanations embedded in questions, actively clarifying questions raised by respondents, and intervening when there are hints of a problem – such as non-verbal communication in a face-to-face interview, pauses in responding and qualified answers. Biemer and Lyberg (2003: 150ff.) argue that more interactive and conversational interviewing by more knowledgeable and better-trained interviewers will lower measurement error, without increasing interviewer effects. High interviewer turnover makes this difficult and so their strategy assumes a more stable workforce, almost exclusively found in central statistical agencies, which offer higher rates of pay and better working conditions than most academic, and especially commercial, survey organizations.

Ideally, interviewing should combine flexibility and standardization, so that for each question interviewers know how much they should assist, and they

have scripted answers to the most common respondent queries. With CATI or CAPI, these answers can be on the interviewer's screen. Interviewing techniques are moving in this direction at the same time as the distinction between self-administered and interviewer surveys is decreasing. Schober and Conrad envisage:

> A next generation of interviewing systems … likely to make use of 'paradata' (process measures) from respondents during the interaction, as a way of diagnosing when help might be needed so that the system can intervene appropriately … One could imagine making use of respondents' typing errors, changed answers, response times, speech disfluencies, or facial expressions to assess their confidence in their answers or their likelihood of having misunderstood. (2008: 6)

Complexity and ambivalence are at the heart of survey interviewing. On the one hand, the idea is to take advantage of the flexibility and intelligence of conversation, so that interviewers select respondents within a household, improve answers by answering respondents' questions, and motivate the respondent. On the other hand, the interviewer's role and the respondent's task completely lack the spontaneity and egalitarianism of conversation.

Since the measurement error resulting from the use of interviewers is proportional to the number of interviews conducted by each interviewer, in order to reduce interviewer effects it is only necessary to lower the average workload. But this results in serious financial and logistic problems. Even if a survey organization is conducting many surveys at one time, so that interviewers can switch between studies, an interviewer who conducts only a small number of interviews with a particular questionnaire makes more errors. So a smaller interviewer effect comes at the cost of greater random error. Also this raises the cost of training, since more interviewers must be trained for each project.

It is unrealistic to expect that interviewer effects can be reduced to zero. Even with careful training, to a small degree, responses may be affected by minor differences in intonation and judgement calls about answering respondent queries, not to mention *respondents'* perceptions of the interviewer. It is cautious and sensible, and the cost is very small, to test for and if necessary statistically control for interviewer effects. To make this possible survey datasets should identify the interviews conducted by each interviewer and provide the interviewers' basic characteristics, including their gender, racialization, any accent, age and experience in the job.

In face-to-face surveys, the assignment of one interviewer to each cluster makes it difficult to separate neighbourhood from interviewer effects, but accounting for the sampling-point clusters still results in correct errors. Dividing geographical clusters between two interviewers in order to obtain a robust measure of interviewer effects is justified only if there is a specific methodological interest or a research focus on neighbourhood effects.

Improving Survey Response

This section takes its name from a recent book by Stoop et al. (2010), titled *Improving Survey Response: Lessons Learned from the European Social Survey*. 'Improving' is apt because there is no silver bullet, only strategies to improve response rates that involve tradeoffs between the cost, representativeness of the achieved sample and the content of surveys. Testifying to the difficulty of the problem, the efforts of a major industry with a financial stake in high response rates have not prevented a steady erosion of response rates in many countries. Where the decline in response rate has been arrested, usually it is because of increased effort to obtain interviews, especially extensive efforts to reach every selected respondent and to change the minds of respondents who refuse initially.

Before considering the research on factors affecting survey response, it is appropriate to review the ideas about why people respond to surveys. Typically, a survey begins with an anonymous request to commit a small amount of time, seldom more than 25 minutes, to complete an interview or mail or Internet survey. Face-to-face interviews are usually not much longer, but they can last an hour or more. If no compensation is offered, the request to answer a survey is comparable to doing a favour for a neighbour or co-worker who is only an acquaintance, making a small donation to charity, or walking a couple of blocks to a voting station. For some respondents, answering a survey is a form of self-expression and perhaps self-affirmation.

For most respondents, none of this is much changed by a small incentive, which acts like an honorarium acknowledging the respondent's contribution, rather than payment for her time. The respondent is more likely to expect an incentive if the survey has an explicitly commercial purpose. Participation in a survey may also involve a straightforward economic transaction; for example it is common practice to pay physicians generously to complete surveys. Longitudinal surveys are a special case because they entail a greater and continuing commitment and because of the expectation that respondents will identify themselves and provide information required to find them if they move. Usually this includes the respondents' name, address and telephone numbers and the names and coordinates of (usually two) people who could find the respondent if direct contact is lost.

Theories of Survey Response

In 1978, Donald Dillman put forward the idea that survey participation involves social exchange, which he differentiates from economic exchange, as follows:

> social exchange is a broader concept. Future obligations are created that are diffuse and unspecified. The nature of the return cannot be bargained over as in economic exchange, but must be left to the discretion of the one who owes it … Fundamentally, then, whether a given behavior occurs is a function of the ratio of … costs and … rewards. (2007: 14)

In return for the time and effort involved in answering the survey, Dillman argues, the respondent is rewarded with the interviewer's 'positive regard', the interviewer's thanks, a sense of accomplishment because he or she responded to a request for help and advice, the chance to 'support group values', and so on (pp. 15–16). He believes that 'providing a tangible incentive, even a token one, is effective because it evokes a sense of reciprocal obligation which can easily be discharged by returning the completed questionnaire' (p. 16). In order to engage this reciprocity, Dillman advocates providing a token incentive *prior* to obtaining a respondent's agreement to participate.

A complete contrast is the application of Ajzen and Fishbein's theory of reasoned action (1980) to survey response by Hox et al. (1995: 5ff.). Instead of invoking a general social principle, they think of survey participation in terms of individual differences. The likelihood of answering a survey is affected, first, by a person's general attitude towards survey research, which affects their attitude towards the topic and the time and effort to complete a specific survey; and, second, by the person's 'normative' beliefs about their friends' willingness to respond to surveys, which affect the person's own views of whether they would respond to surveys on different topics. In testing the model in an introductory psychology class, measures of attitudes and norms were disappointingly weak predictors of survey response (p. 8). Also, the theory has a certain tautological character when it predicts that people who approve of a survey are more likely to respond.

At this time, the closest to a conventional explanation of survey response is Groves et al.'s (2000) 'leverage-saliency theory'. The idea is that individuals vary in the importance they ascribe to the different attributes of a survey, including the survey topic, any incentive and the credibility and worth of the survey sponsor. In deciding whether to answer a survey, the respondent combines her assessments of its attributes, weighting each attribute according to its salience.[7] Someone who is concerned about public policy, for example, might be convinced by an appeal emphasizing the impact of the findings, while being unaffected by a monetary incentive; while someone with no interest in the survey topic might be attracted by the incentive. Groves et al. (2000: 305ff.) demonstrate that respondents who are not interested in the survey topic are most affected by incentives, while Groves and other colleagues (2004) observe a similar, though weak, effect.[8]

[7]More precisely, the predictor of response is the sum, over all the attributes of a survey, of the *product* of the importance of each attribute and the evaluation of that attribute.

[8]In a comprehensive review of the effect of incentives on telephone survey response, Cantor et al. (2008) find no pattern of variation in the effect of incentives on demographic groups, except that 'respondents with intrinsically low-response propensity are more affected by an incentive' (p. 497). While they interpret this as inconsistent with the prediction of leverage-saliency theory, the theory predicts that incentives raise the response rates of individuals who are not otherwise motivated to cooperate, not that particular socio-demographic groups are differentially sensitive to incentives.

Exchange and leverage-saliency theory are rational choice theories, as is Singer's (2011) very similar 'benefit–cost' theory, whereby respondents agree to be interviewed if the benefit outweighs the costs. Dillman cautions that social exchange is inexact and, compared to Groves et al., he gives more emphasis to the benefit of a respondent's opportunity to express his or her mind and gain the attention and respect of the interviewer.

According to leverage-saliency theory, respondents decide whether to answer a survey by weighing the pros and cons; indeed they use an illustration of a balance scale. Lin and Schaeffer (1995) observe that this model implies that individuals are located along a 'continuum of resistance', ranging from 'high-propensity' respondents who are most easily contacted and persuaded, to the 'lowest-propensity' respondents who are hard to find, not interested in the topic and indifferent to an incentive. The response propensity can be estimated empirically from the number of calls required to reach a respondent and how readily he or she consented to an interview. It is possible to use propensity scores to weight the data, giving higher weights to respondents with low propensity, because they 'represent' respondents who mostly refused to answer the survey. We return to this point in the last section of this chapter.

Neither exchange theory nor leverage-saliency theory effectively addresses the decline in survey response rates. In focusing on a transaction between the respondent and the survey sponsor, exchange theory does not connect survey participation to characteristics of the individual or society. While leverage-saliency theory links respondents' values to survey participation, it is hard to argue that, say, declining response rates reflect a decreased interest in public policy. To a significant degree, the decline in survey response must reflect changing attitudes towards surveys themselves and towards the institutions – governmental, academic and commercial – that sponsor and conduct them.

Field Strategies for Increasing Survey Response

In empirical research on survey response spanning more than 40 years, from a number of countries and using different modes with widely varying topics, there is a high degree of consensus on the factors affecting survey response. These can be divided into three classes: attributes of an entire survey; data collection procedures; and the characteristics of individual respondents. This section focuses on aspects of survey data collection practices that the researcher does control: efforts to contact and convince a respondent to participate; the pitch made to respondents; and incentives.

The 'fixed' attributes of a survey that affect the response rate include the identity of the survey sponsor and data collection agency, the survey mode, the topic of the survey and the target population. Surveys conducted for government agencies have higher response rates than academic surveys, which have higher response rates than commercial surveys; and there are corresponding differences in the response rates of surveys conducted *by* government agencies and academic and

commercial organizations. Also, longer surveys and surveys on more difficult topics have lower response rates. On these points, see meta-analyses by Heberlein and Baumgartner (1978), Goyder (1987) and Hox and de Leeuw (1994). These findings are of no practical help in increasing survey response rates because, for a given survey, they are not subject to change. Similarly, the demographic, socio-economic and attitudinal characteristics of respondents affecting the response rate are dictated by the composition of the sample. Nice reviews can be found in Groves and Couper (1998), Stoop (2005: 64ff.) and Stoop et al. (2010: 24ff.).

Persistence The most important predictor of response that is under the control of the data collector is simply persistence, which involves no more than repeated and effective attempts to contact respondents. For telephone surveys it is common to specify a minimum number of calls to each selected number, to space the calls over a long period (typically at least three weeks), and to vary the time of day and day of the week when calls are made. The idea is to contact respondents who are not often at home, who do shift work or have unusual hours for another reason, or who go on vacation. With each unanswered call the probability of ever obtaining a response declines and analysis of the success of reaching respondents with different numbers of calls indicates when it makes sense to give up. This also holds for face-to-face surveys, though cost considerations loom much larger.

For interviewer surveys it is important to systematically follow up successful contacts that do not result in an interview, including respondents who repeatedly say they are busy. Persistence also involves attempting to 'convert' refusals into completed interviews, unless the initial refusal was adamant or the jurisdiction does not allow a subsequent call. Depending on the circumstances, 'converted refusals' typically account for 15–30 per cent of completed interviews. Refusal conversion is highly cost effective; not only is there information about how to contact the person. A relatively high proportion of calls are successful.

Similar considerations hold for self-administered surveys. For mailed surveys, where the cost of each contact attempt is significant, the normal routine is to send an initial questionnaire, followed by a brief reminder – often just a postcard, followed by a second copy of the questionnaire, followed by a second reminder. Accounting for the vagaries of mail distribution, these might be sent at 10-day intervals. A typical distribution of the percentage of the total returns after the four contacts might be 40 per cent after the first message and questionnaire, then 25, 20 and 15 per cent for the three subsequent mailings.

For Internet surveys the cost of additional email solicitations is very low or negligible and the response to each is nearly immediate. After each email request there is a sharp peak of responses in the first day or two, followed by a steep decline, so that there are very few additional respondents after the fourth or fifth day. There is significant additional response until at least five or six messages.

When the contact information is available, another strategy is to employ a second mode of contact, such as making telephone calls to non-respondents of face-to-face,

mail or Internet surveys or sending postal mail reminders to telephone or Internet survey non-respondents. The novelty of a reminder sent in a different mode and perhaps the implicit message that the response is important add to its value. Special delivery mail or other means of embellishing an invitation can increase the response rate. In some countries initial recruitment by telephone is effective for face-to-face interviews and results in significant cost savings because the interviewer can make an appointment (Stoop et al., 2010: 131ff.).

For telephone surveys, a common strategy is to send an 'advance letter' describing the survey and saying the household will receive a call. De Leeuw et al. (2007) found an increase of 5–10 per cent in response rates, which is substantial and cost effective, but no such effect was found by Singer et al. (2000) or Holbrook et al. (2008).

The Pitch While care should be taken in composing the appeal to respondents, what can sensibly and honestly be said about a given survey is quite limited and, without significant information about the respondents, mail and Internet survey invitations cannot be 'tailored' to individuals. Writing these invitations is a form of advertising on which many textbooks provide advice. In a nutshell, the appeal should be brief (three or four short paragraphs), engaging and direct, invoke a good cause, attest to the confidentiality of responses, and where appropriate note the approval or authorization of significant others or organizations. For advice on 'cover letters', as well as reminders for subsequent contacts see Dillman (2007: 149ff.).

Surveys with interviews are different because there is potentially some exchange between the interviewer and selected respondent and there is substantial variation in interviewers' abilities to recruit respondents (Blom et al., 2011). Effective interviewers steer the conversation in the direction that responds to subtle clues from the respondent and they know to break off the conversation and promise to call back when a respondent is on the verge of refusal (Groves and Couper, 1996; Snijkers et al., 1999; Schaeffer et al., 2010: 445ff.).

Incentives There is no reason to dispute the conclusions of Church's 1993 meta-analysis of studies of controlled experiments which found that: cash incentives, and to a smaller degree gifts and lottery tickets, provided *before* a survey rather than after, increased mail survey response rates; that the effect of incentives increased with their size; and that promised incentives paid only after a survey was completed had small or negligible effects. On average, providing a cash incentive prior to the survey increased the response rate by 19 per cent, but the increase was only 4.5 per cent when the incentive was paid only after the survey was completed. The median cash incentive of the surveys analysed by Church was just $0.86 – even at the time a very small amount. Non-monetary incentives were much less effective. They increased the response rate by 8 per cent if offered in advance, by just 1 per cent afterwards.

Singer et al. (1999) addressed the same questions about telephone and face-to-face surveys, again with a meta-analysis of experimental comparisons. While the surveys they considered were more recent than those examined

by Church (51 per cent were in the 1990s) and the mean payment was much higher (US$11.39), the results were similar. Prepaid cash incentives increased response rates to a greater extent than incentives paid after completing the survey, and gifts had smaller effects than cash. Incentives paid after the survey still increased response rates significantly, but the overall effect of incentives was weaker than Church reported.[9] Perhaps an interviewer's promise of an incentive is more credible than a written promise to pay. They found positive effects of incentives for both cross-sectional and panel surveys and for high-burden (longer or more difficult) and low-burden surveys. Singer et al. (2000) found that a US$10 incentive increased telephone survey response rates by over 10 per cent and that being paid to complete a survey did not decrease a respondent's willingness to answer a subsequent survey with no incentive. Promising payment after the interview had no beneficial effect at all.

In a more recent meta-analysis for RDD telephone surveys, most of which were conducted after 2000, Cantor et al. (2008) found that a prepaid US$1 or $2 incentive increased the response rate for a screening interview (where a household member is contacted and a respondent selected within the household) by about 5 per cent, and a $5 incentive increased it by about 8 per cent. An incentive promised after the interview also increased response rates, but only if the amount was at least $15.

Unfortunately, almost all the reported research on incentives is American. Stoop et al. (2010: 102) note that the different national components of the ESS each adopt their own policy on incentives and that the response rate was 3 per cent *lower* in countries which provided incentives than those that did not. Hopefully, this is because researchers are more likely to offer incentives in countries where response rates are lower.

Incentives pay for themselves when the value of the additional responses is greater than their cost. But, typically, most respondents would do a survey without an incentive, even if the incentive induces more response. Ideally one would pay only respondents who would not respond otherwise, but rewarding uncooperative behaviour in this way might not be ethical and risks poisoning the well for future surveys.

Response Rates, Bias and Error at the Margin

While the response rate is widely considered the best indicator of survey quality, the harder question is whether there are any differences between surveys with, say,

[9]Singer et al. (2000: 225) describe the effects as 'relatively modest once other variables have been controlled'. Because the effects are reported only as standardized regression coefficients, so it is not possible to describe them in terms of percentages. They also find that incentives have a smaller effect on surveys that have a high response rate without an incentive – an effect that might disappear if they predicted the logits rather than the absolute response rates.

45 and 50 per cent response rates, when the 5 per cent increment is obtained with incentives, more effective refusal conversion, or making extraordinary numbers of calls to reduce non-contact.

The preponderance of evidence is that a small increase in the response rate does little to change population estimates (Curtin et al., 2000; 2005; Groves, 2006; Keeter et al., 2000). While this is a sensible test if the goal is only to estimate population characteristics, it is a weak statistical criterion, because that 5 per cent of difficult-to-reach respondents are a small proportion of all respondents and their presence would measurably change the population figures only if they were radically different from the other respondents. In any event, the evidence is that hard-to-reach and reluctant respondents are *not* very different from the easier-to-reach respondents, and that both may be different from the remaining respondents.

The diminishing returns of efforts to increase the response rates by a few per cent suggests that *for a given survey* the target population consists of two classes. One class includes potential respondents, differentiated according to how easily they can be reached and convinced to participate, while the second consists of persons who will never participate. That second class includes people who were never contacted despite repeated attempts and people who were contacted but refused to do the survey – who immediately refused categorically or who give a 'soft' refusal at first, but did not respond to subsequent appeals.

Fricker and Tourangeau (2010) suggest that hard-to-get respondents contribute disproportionately to measurement error because their lack of interest is compounded by an inability or unwillingness to respond *well*. Their analysis of the (US) Current Population Survey and the American Time Use Surveys revealed the predicted negative relationship between response propensity and data quality, but it was very weak and other research also finds little or no relationship (for example, Yan et al., 2004).

Using the third wave of the ESS panel surveys for Belgium, the Netherlands, Norway and Sweden, Kaminska et al. (2010) look at the relationship between propensity to respond and data quality. They find that respondents with lower cognitive skills are more likely to satisfice, measured by item non-response, straight-line response (giving the same answer to a whole series of questions), choosing the middle response, and giving inconsistent answers to substantively similar questions. Their measures of cognitive skills included education, interviewer evaluations of respondents' understanding of the survey questions, and how often respondents asked for clarification. They find a negative relationship between response propensity and data quality, *but* controlling on cognitive skills makes the relationship disappear. Kaminska et al. argue for the need to actively recruit low-propensity respondents because they are disproportionately from groups with lower cognitive ability, which tend to be under-represented in surveys. The implication is that recruiting reluctant respondents guards against bias in comparisons of population subgroups, even if it does not affect estimates for the total population.

If the members of a survey sample can be characterized by a reasonably smooth distribution of the propensity to respond, it should be possible to compensate for

non-response by weighting that gives more influence to *low*-propensity respondents. For respondents, the propensity can be measured by the effort required to obtain a response, indicated by the number of calls required to obtain an interview and distinguishing initial respondents from 'converted' refusals. But Stoop et al. (2010: 264ff.) find that propensity-based weighting produces almost no difference in variable distributions and from this conclude that respondents do not fall along a single 'continuum of resistance'. This is further support for the 'two-class' argument that differentiates potential respondents. The size of the second class is a function of the survey topic and length and the national setting, and it has increased over time.

Now, say that time, mode, cost or other constraints result in a response rate of 30 per cent for a population and on a topic where an exceptional and expensive survey could conceivably achieve a 70 per cent response rate. So, many potential respondents are not successfully surveyed. Although propensity-based weighting would still not tell us anything about the 30 per cent absolute non-respondents, it might produce better estimates for the 40 per cent of the population who are potential, but not actually surveyed, respondents. A nice example is provided by Biemer and Link (2008).

Beyond what is known from the sampling frame, it is difficult to find out much about non-respondents in telephone and self-administered surveys. In face-to-face surveys, however, it is possible to gather information about a dwelling and its neighbourhood and possibly fragmentary information about the selected respondent with a brief questionnaire ('If you won't do the survey, would you consider answering a few short questions?'). Nice examples of this kind of research implemented as part of the ESS are described by Stoop et al. (2010: 243ff.).

Statisticians have also considered the relationship between non-response and bias, which is different for each variable in a survey, as the formulation due to Bethlehem (1988; 2002: 276) shows. If ρ is the propensity (just the probability) that a person will answer a survey and Y is any variable measured in the survey, then the bias in $\bar{y}$, the estimated mean of Y, is $\text{cov}(\rho, Y)/\bar{\rho}$, where $\text{cov}(\rho, Y)$ is the covariance of ρ and Y, and $\bar{\rho}$ is the response rate. For a given survey, $\bar{\rho}$ is fixed, so the degree of bias depends on the relationship between the variable Y and the probability that a person will respond to the survey.[10] There is bias in estimates of charitable giving, recall, because the donors are more likely to answer a survey about donations! Similarly, more satisfied students and employees are more likely

[10]The formula allows us to remove the bias in estimates of $\bar{y}$ *if* $\text{cov}(\rho, Y)$ *is known*. But that covariance must be estimated from the respondents only, since Y is missing for the non-respondents. For respondents, ρ can be estimated (with some error) from the number of contracts required to obtain a completed survey and, for interview surveys, also whether the respondent refused before being successfully interviewed. If the survey has contacted almost all the selected respondents, then the value of ρ for non-respondents is close to or exactly zero and usually it is not possible to differentiate among them. The question is whether the value of $\text{cov}(\rho, Y)$ *in the realized sample* is a good estimate of the *population* value of $\text{cov}(\rho, Y)$.

to respond to surveys conducted by their school or employer, and surveys about attitudes towards surveys are highly prone to bias.[11]

Groves and Peytcheva (2008) were able to locate published reports of 59 studies with 959 estimates of non-response bias, and they conclude that

> the meta-analysis shows much variability in nonresponse bias within surveys, across estimates … when influences on survey participation are themselves measured in the survey, they will show the largest nonresponse bias. To predict what survey estimates are most susceptible to nonresponse bias, we need to understand how each survey variable relates to causes of survey participation. (p. 183)

Overall, higher response rates result in lower bias, but some surveys with a low response rate have very little bias. This points to the need to determine whether the data collection process more effectively reaches individuals who are distinctive in terms of the particular survey measures. If there is such a relationship, there is a strong argument for weighting to compensate for bias, ideally using auxiliary measures from the sampling frame, but otherwise using response propensity estimates derived from the survey data.

A disturbing and important finding from Groves and Peytcheva's study is that when the overall mean of a variable is biased, subgroup differences, for example comparing women and men or age groups, also tend to be biased (2008: 182). If simple comparisons are biased, it is likely that model coefficients are also biased.

Conclusion

Survey data collection is the subject of a huge body of publication, far from completely summarized in this chapter. Research in this area largely involves weakly theorized comparisons of alternative strategies, driven by the hope of gathering better or more data without increased cost. Given a survey topic and population and the funds available, there is an impressive body of practical knowledge on how best to gather data and on the key tradeoffs, for example between response rate and survey length and between the surveying of more respondents (to lower sampling error) and a higher response rate (to decrease bias). There are practical answers questions such as how many attempts should be made to contact respondents and how the survey length affects the response rate. Even when the

[11]One strategy for studying the effect of attitudes towards surveys on survey response is to begin with a captive audience that can initially be surveyed in its entirety, for example students in a university class, and then regress whether members of that group complete a subsequent survey on their initial attitudes. But it is difficult to find a captive group that is representative of a typical non-institutional survey population and it is hazardous to generalize about population phenomena from captive groups, such as college students.

reported evidence is mixed, for example regarding whether 'advance letters' to selected respondents increase the response rate, often it is possible to predict the outcome for a particular topic and sample. The people who manage surveys have these answers at their fingertips.

Both for the operational decisions in conducting surveys and the academic understanding of the field, the methodological research has done more to assess the effects of alternative strategies on data quality than to identify superior methods, at least when the cost is fixed. Because of logistic and cost constraints, only rarely do 'big' questions – such as whether to conduct a telephone or face-to-face survey, or use a longitudinal survey or repeated cross-sectional surveys – identify realistic alternatives. At the same time, the well-developed knowledge of the impact of survey mode on data quality represents a major research achievement of increasing practical use as mixed-mode surveys proliferate. Similarly, while the reported magnitude of interviewer effects is highly variable and (except for sensitive questions) cannot be predicted accurately from the survey topic or sample, we know both how to analyze them and that the potential effects should not be ignored.

Advances in survey data collection have been driven by the increasing sophistication of communications and computing and by researchers' demands for higher-quality and more complex surveys. Not only experiments, but also the questions routinely used to ask about employment, education and other conventional demographic characteristics routinely employ complex 'skips' that demand computerized questionnaires. The second major source of innovation has been the changes in communications technology, especially how telephones are used, and increasingly negative public attitudes towards surveys in general and towards the organizations that sponsor them and collect data. Especially in the last decade, the struggle has been to prevent a decline in the capacity to gather data.

The entire enterprise of survey data collection has been shaped by the idea that survey participation involves a worthwhile and voluntary activity, and the task of data collection is to spread that message and accommodate the volunteers. Even when there are 'incentives', which are not the norm in academic, government or even commercial surveys, the idea is to offer respondents a token of a survey's worth rather than to buy their time. It is very difficult, however, to imagine the alternative of a market for survey participation of the general population. The inherent problem of voluntary participation, however, is that it is vulnerable to changes in the capabilities and credibility of communications technology and in respondents' willingness to answer.