

A Good Mix? Mixed Mode Data Collection and Cross-national Surveys

Peter Martin

The Anna Freud Centre, London

Can cross-national surveys benefit from mixed mode data collection? This article provides a classification of the different ways in which modes of data collection may be mixed within a cross-national survey, and investigates the methodological consequences of such designs. Mixed mode designs have the potential to lower survey costs relative to single-mode face-to-face surveys, while maintaining higher response rates than cheaper modes alone could. Yet since responses to survey questions are not always independent of the survey mode, mixed mode designs endanger cross-national measurement equivalence (as well as, in the case of time series surveys, diachronic equivalence), so that cross-national comparisons (and analyses of change over time) lose internal validity. These problems can be mitigated by careful questionnaire and survey design, but won't be entirely overcome in many cases. The use of mixed mode designs in cross-national surveys therefore needs to be accompanied by methodological research to establish the likely consequences for measurement.

Key words: survey methodology; mixed mode surveys; comparative research; cross-cultural research.

Acknowledgements

I would like to thank my colleagues at the Centre for Comparative Social Surveys at City University, London, for valuable comments. A previous version of this article has been published as CCSS Working Paper No. 02 „What makes a good mix? Chances and Challenges of Mixed Mode Data Collection in the European Social Survey.“ Available online at: www.europeansocialsurvey.org[Access 6/12/2011].

Can cross-national surveys benefit from mixed mode data collection? Researchers hope that mixed mode surveys may be more cost efficient than single-mode surveys, and that they may assist in efforts to obtain data from difficult-to-reach respondents. However, mixed mode data collection may lead to measurement error that is often difficult to predict or quantify. In cross-national surveys, this problem is particularly serious, since in this context, “mixed mode” most likely means that different countries will use different modes, or combinations of modes. The associated mode effects risk undermining the principle of equivalent measurement, which a cross-national survey must satisfy if its data are to be useful for comparative cross-national research.

A number of useful overviews over the potential and the problems of mixed mode data collection have been published (de Leeuw, Dillman and Hox 2008: 299; de Leeuw 2005: 233; Dillman, Smyth and Christian 2009). Yet to the best of my knowledge, we do not yet have a systematic reflection on what mixed mode data collection designs would imply for cross-national surveys. It is an important topic, as the co-ordinators of cross-national surveys face the twin problems of rising survey costs and dwindling response rates in many countries. One of the best-known cross-national time series surveys, the World Values Survey, is in fact already a mixed mode survey, although this is rarely made explicit (World Values Survey 2009).

The present article provides a classification of the different ways in which modes of data collection may be mixed within a cross-national survey, and investigates the methodological consequences of such designs. The empirical evidence on which we can draw to gauge the feasibility of mixed mode surveys in an international context is thinner than desirable. Yet some excellent studies have been done, which will be summarized here. A particular emphasis will be given to the European Social Survey’s Mixed Mode Methodology Programme, the only research programme to date that has focused specifically on mixed mode designs in cross-national surveys.¹

INTRODUCTION

Some say that mixing modes of data collection is the future of survey research. At a time when survey data collection costs are rising and response rates falling, survey researchers are confronted with ever greater practical difficulties in their endeavour to collect representative data from populations. Mixed mode data collection is sometimes hailed as a possible contribution to overcoming to these problems, as mixing modes may be part of a strategy that allows researchers to optimize the expenditure of effort and resources in respondent recruitment. If responses from easy-to-reach and motivated sample members can be collected

in relatively inexpensive data collection modes, a greater proportion of overall resources would be available for chasing hard-to-reach and unwilling sample members. A famous example for a successful implementation of such a strategy is the American Community Survey (US Census Bureau 2009), which employs three modes (postal, telephone, and face-to-face) sequentially in order to gather as many data as possible by cheaper modes, before resorting to the deployment of interviewers to target persistent non-respondents.

However, mixed mode surveys also create problems for survey research. Apart from the practical challenges of preparing fieldwork documents for several survey modes, and of multi-mode sample management, the use of different modes for different respondents highlights the influence that mode of data collection has on survey measurement. Simply put: with some survey questions, the answer a respondent will give will be different depending on the mode of data collection. Different modes cause different measurement error, and different biases. This would be a problem for any survey where respondents are not allocated to modes randomly, because insofar as different population subgroups differ in their propensity to select one mode rather than another, subgroup comparisons on variables that are subject to mode effects would be biased. Yet the problem is even more serious in cross-national surveys.

Cross-national surveys depend on the principle of equivalent measurement (Jowell 1998: 168). It is the very rationale of their existence: to provide rigorous measures of variables in all participating countries in an equivalent way, so that comparisons between countries are not confounded by country-specific measurement error. High-quality cross-national surveys, such as the European Social Survey, invest a lot of effort into the promotion of measurement equivalence, striving for example to harmonize questionnaires, to produce accurate translations of survey questions across languages, and to develop principles of sampling that are binding for all countries (Fitzgerald and Jowell 2011).

Given these efforts to maximize cross-national equivalence, mixing modes is a hazardous undertaking. If different countries employ different modes of data collection, mode effects would undermine the principle of equivalence, as they can result in biases that would, at the stage of data analysis, confound comparisons between countries. Yet external pressures to mix modes across countries are strong: different countries have different “survey climates” (Jowell, Kaase, Fitzgerald and Eva 2007: 1) – a summary term that refers to the conditions under which surveys are conducted in a country, including conditions that may be more favourable to data collection in one mode rather than the other. Such conditions include interviewer costs, the expertise of survey fieldwork agencies, respondents’ experiences with and expectations of survey agencies, and mode coverage (such as internet penetration and rates of telephone ownership).

It is time, then, to consider in a systematic way the methodological challenges mixed mode data collection would face in the context of cross-national surveys. In the next section, I shall introduce a few basic definitions. The World Values Survey will serve as an example that illustrates one type of cross-national mixed mode survey. I will then go on to discuss threats to measurement equivalence that mixed mode data collection entails. The evidence points to the existence of substantial mode effects that would bias cross-national and diachronic analyses – but it is not clear that, compared to other measurement errors that surveys routinely suffer from, mode effects are sufficiently large to rule out mixed mode data collection for cross-national surveys altogether. Finally, I will argue that any cross-national survey that values equivalent measurement needs to accompany the introduction of mixed modes with methodological experimentation to gauge the extent of bias introduced by mode effects.

MIXING MODES IN INTERNATIONAL SURVEYS: SOME DEFINITIONS

The term “mixed mode survey” needs to be clarified, and we need to introduce a number of important distinctions. De Leeuw (2005: 233) and Groves et al. (2004: 163–165) have provided excellent overviews of the issues around mixed mode surveys in general, and the following discussion is indebted to their work.

Multiple modes survey systems. First, it is important to realize that many surveys are in fact “multiple mode survey systems” (cf. de Leeuw 2005: 237), even if they employ only a single mode of data *collection*. Thus, respondents may be first contacted and possibly even selected in one mode, but have their answers collected in another. In a cross-national survey, countries may of course differ in the way they use such practices. However, as long as the data are all *collected* in the same mode, this practice has no known impact on measurement error (de Leeuw 2005: 235), and as such does not seem to pose a threat to measurement equivalence.

Multiple modes data collection. I shall also distinguish between *mixed* mode data collection and *multiple* modes data collection. The terms are sometimes used interchangeably, but it is useful to keep them apart. For the purpose of this paper, a “multiple mode design” is defined as a design whereby data are collected using more than one mode, but any survey question is posed in the same mode to all respondents. Put differently: in a multiple mode design, mode varies by survey item, but not by respondent. A typical example of a multiple mode design is a self-completion section within a face-to-face interview – a device sometimes employed with the aim of allowing respondents to answer sensitive questions without disclosure to the interviewer. A multiple mode design poses no known problems for cross-national and longitudinal comparisons, if it is implemented in the same way in all countries, and if it remains the same across survey waves.

Mixed modes data collection. In contrast, mixed mode data collection signifies a design whereby the mode of survey administration is allowed to vary by respondent. That is, different respondents answer the same survey questions in different modes. There are two common ways in which such a design is implemented: in a **concurrent design**, respondents are given a choice of mode (that is, several modes are offered concurrently); in a **sequential design**, sample members are first invited to complete the survey in one mode (typically, but not necessarily, the cheapest one), and subsequent modes are only offered to those who fail to respond to the earlier invitations.

An example of a (single-country) sequential mixed mode design is the American Community Survey, already mentioned above, which uses postal questionnaires, telephone interviews, and face-to-face interviews. Sample members are initially invited via postal questionnaires that are sent to their homes. Non-respondents are then followed up by telephone calls, where telephone numbers are available. Finally, all remaining non-respondents are visited by interviewers at their homes. This procedure increases response rates relative to a single-mode postal survey (US Census Bureau 2009).

As should be clear by now, this paper focuses on mixed mode data collection, rather than on multiple modes of data collection or multiple modes survey systems. The next step is to acknowledge that in cross-national surveys, and more particularly in cross-national time series, mixed mode data collection can take different forms. I would like to introduce the distinction between three different, albeit not mutually exclusive, mixed mode designs, which I would like to call, respectively, Across-country Mixed Mode (ACMM), Within-country Mixed Mode (WCMM), and Across-time Mixed Mode (ATMM).

I. Across-country mixed mode data collection (ACMM). In a cross-national survey, different countries may use different modes of data collection. This approach is taken in a number of international surveys, such as the Pew Global Attitudes Project (Pew Global Attitudes Project 2009), where data are collected by face-to-face interview in some countries, and by telephone in others; and the World Values Survey, where most countries collect data by face-to-face interview, but some use self-completion questionnaires (World Values Survey 2009). Across-country mixed mode data collection, or ACMM, in this sense, can occur even if each country uses only one mode. The cross-national survey as a whole becomes a mixed mode survey as soon as at least two countries differ in the mode of data collection.

II. Within-country mixed mode data collection (WCMM). A special case of ACMM occurs if one or several countries collect data with more than one mode, using different modes to survey different respondents. This is what I would like to call Within-Country Mixed Mode (WCMM). Methodologically, WCMM brings in

additional complexity: the fieldwork agency in a WCMM country has to administer different modes of data collection, and to oversee a more complex survey system, because questionnaires have to be produced in two or more formats, different respondents might be contacted, followed up, and surveyed in different ways, and data from different input templates have to be combined. When analysing data collected by WCMM, mode effects may affect comparisons between countries that use different designs (either single-mode versus mixed-mode, or different mixed-mode designs). In addition, analyses involving subgroups from a mixed-mode country may be affected by mode effects, if the process of selection into different modes is non-random (as it is likely to be).

III. Across-time mixed mode data collection (ATMM). In a cross-national time series, it is also possible that one or several countries change from one mode to another between waves of data collection, or even change from a single-mode design to a WCMM design (or vice versa). Any such change would render the survey an across-time mixed mode survey (ATMM). An example of a single-country ATMM design are the British cohort studies, two longitudinal surveys following birth cohorts of British residents from birth throughout the life course.² They started out as face-to-face only surveys, but have since adopted an ATMM design, alternating face-to-face and telephone interviews in consecutive waves. For the researcher interested in longitudinal analysis, an ATMM design means that real change over time in variables of interest may potentially be confounded with measurement effects due to different modes of data collection.

The three mixed mode designs – ACMM, WCMM, and ATMM – are not mutually exclusive. In fact, WCMM of even a single country within a cross-national time series implies that the survey is also ACMM. Similarly, ATMM implies ACMM, unless all countries make the same change at the same time. On the other hand, it is possible to have an ACMM or ATMM design without WCMM: namely, if every country uses only one mode of data collection, but some use different modes from others.

THE WORLD VALUES SURVEY AS A MIXED MODE CROSS-NATIONAL TIME SERIES

The concepts developed in the previous section can be illustrated through a look at the World Values Survey (WVS). Table 1 shows the modes of data collection of five WVS waves, in five countries. We see that the WVS is a mixed mode survey, both in the sense of ACMM and ATMM, although not WCMM. Altogether, three modes of data collection have been employed: face-to-face interviews, postal questionnaires, and web-based questionnaires. Looking diachronically at each country separately, two time series are single-mode: New Zealand has conducted all its waves using postal questionnaires, and the United Kingdom has used

exclusively face-to-face interviews. However, the time series of the three other countries are ATMM. Each of these three began using face-to-face interviews, but then changed to a different mode: Australia changed to postal questionnaires in the third wave, and Japan in the fourth wave – while the United States adopted web questionnaires in the fifth wave, collecting data from a pre-established panel of respondents, rather than drawing a fresh probability sample.

Table 1. Modes of data collection in five countries of the World Values Survey, 1981–2008

Wave number	1	2	3	4	5
Year	1981–1984	1989–1993	1994–1999	1999–2004	2005–2008
Australia	Face-to-face	[no survey]	postal	[no survey]	postal
Japan	Face-to-face	Face-to-face	Face-to-face	postal	postal
New Zealand	[no survey]	[no survey]	postal	[no survey]	postal
United Kingdom	Face-to-face	Face-to-face	Face-to-face	Face-to-face	Face-to-face
United States	Face-to-face	Face-to-face	Face-to-face	Face-to-face	Web (panel)

Note: The five waves of the WVS included between 21 and 71 participating countries. Most of these countries have conducted all their surveys using face-to-face interviews. The five countries shown here were selected for the purpose of illustration. While cross-national comparisons and trend analyses may be affected by mode differences in comparisons involving the countries shown here, this is not the case for analyses that draw on face-to-face countries and waves only.

If we consider the five waves of the WVS, with respect to the five countries shown, we note that the first two are single-mode waves, employing face-to-face data collection only. The three subsequent waves, however, are mixed mode (ACMM) waves, as Waves 3 & 4 employ face-to-face interviews in some countries, and postal questionnaires in others. In Wave 5, the use of web questionnaires in the US means that a third mode is added.

As the following discussion will show, the mixed mode design of the WVS risks introducing biases into any analysis that involves the five countries presented in Table 1 in cross-national comparisons, or the investigation of change over time. The possibility of such analyses is of course the very rationale for conducting a cross-national time series survey.

MEASUREMENT EQUIVALENCE

Many threats to equivalence exist in comparative research. In order to claim equivalence of measurement, a number of assumptions need to be made, such as equivalence of translation, equivalence of population coverage and sampling,

equivalence of non-response bias, and equivalence of the meaning of concepts across national and cultural boundaries. Most of these cannot, in practice, be proven scientifically for every survey and every survey question. Survey designers rely on a mixture of evidence, experience and common sense to argue for the plausibility of assuming equivalence, and sometimes (as in the case of the possibility of non-response bias), when all affordable efforts have been made, the absence of bias is tacitly assumed, often without evidence to either support or undermine the assumption.

The drawback of introducing mixed mode data collection into a cross-national survey is that this adds another threat to equivalent measurement. Some effects of mode on measurement are relatively well researched, such as the differing propensity of different modes to engender socially desirable responses, and the different response tendencies associated with visual and aural presentation of response options (primacy and recency effects). Other types of mode differences remain unexplained. What is more, most research on mode differences has been conducted in North America and Western Europe, and it is not necessarily clear that mode differences will be culturally invariant, adding a further layer of complication. Neither is it clear that mode differences will be invariant over time (Holbrook, Green and Krosnick 2003: 111).

THREATS TO EQUIVALENCE

Four types of equivalence threats due to mixed mode designs have been identified in the scientific literature: mode effects on item measurement; mode-specific sampling frame coverage; design-specific selection effects; and the practical challenges of running a multi-mode survey. These issues already complicate mixed mode data collection in a single nation survey, but they are in many ways intensified in the situation of a cross-national survey. As we discuss the threats to equivalence in turn, it will be important to keep in mind the distinctions between ACMM, ATMM and WCMM. Because face-to-face interviews are by far the most frequent mode employed in scientific cross-national surveys, and because they usually yield the best data quality in terms of response rate, and, sometimes, response accuracy, we will in the following discussion discuss other modes in comparison to face-to-face interviews.

Measurement effects and data quality

Measurement effects are biases introduced through the influence of the measurement instrument on the measurement itself. Insofar as we are interested in measuring parameters as precisely as possible, we must be concerned with systematic measurement error, with systematic deviations from the true score that are due to

data collection mode. However, insofar as we are interested in cross-national and longitudinal equivalence, we must also be concerned with the different effects of different modes. Maximizing measurement precision in only some, but not all countries, might jeopardize equivalence. For example, a more precise measurement in one country (say, by employing a self-completion mode that has a smaller social desirability bias than other modes on a given variable) may not be advantageous if measurement equivalence with other countries (that use different modes) is thereby lost.

Careful questionnaire development is pivotal to reducing mode differences in measurement. Dillman, Smyth and Christian (2009) have laid out principles for a “unified mode design” that aims to make the question stimulus as similar as possible across modes. Some of the largest measurement differences due to mode found in methodological research have resulted from differences in question design, such as the use of different response categories in different modes, or differences in layouts of questionnaires. Yet even the most careful questionnaire design cannot eliminate all mode differences, and it is to those recalcitrant mode differences that we now turn.

Telephone interviews

Telephone interviews have been shown to be subject to measurement effects relative to face-to-face interviews in two respects: satisficing and social desirability. There is consistent evidence that telephone interviews tend to elicit more socially desirable responses than face-to-face interviews. This has been established in the USA, where Holbrook et al. (2003: 79) found evidence that telephone respondents give more socially desirable responses than face-to-face respondents across three national studies on a variety of attitudinal and behavioural variables with empirically established social desirability connotations, including political interest, attitudes to government aid for African Americans, voting participation, and attendance at religious services.

Research conducted within the ESS Mixed Mode Methodology Programme (Jäckle, Roberts and Lynn 2006; Jäckle, Roberts and Lynn 2010: 3) confirmed the greater propensity of telephone interviews to elicit socially desirable responses in an experiment conducted in Hungary. The results of the latter experiment also suggested that it was the location of the interviewer (i.e., on the phone and not in the respondent’s home) rather than the absence of showcards (which are ordinarily used in many ESS questions) that caused the social desirability bias in the telephone interviews. Social desirability effects were found with respect to questions concerning political interest, political efficacy, attitudes to immigration, attitudes to household division of labour, religiosity, attendance of religious services, time spent watching television, and household income.

Telephone interviews have also been suspected of yielding lower quality data than face-to-face interviews (Roberts 2007). In many studies, however, the evidence is mixed: some indicators of satisficing suggest that respondents satisfice more in telephone surveys, while others suggest the opposite, or no difference. For example, in an ESS experiment (Roberts, Eva, Lynn and Johnson 2008), 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. The strongest mode effects were found in Switzerland, while hardly any were found in Hungary. Overall, it is not clear whether telephone interviews necessarily lead to lower quality data than face-to-face interviews.

An MTMM experiment, conducted as part of a larger WCMM study in the Netherlands using the ESS questionnaire, found that telephone interviews yielded the lowest data quality relative both to face-to-face interviews and to web questionnaires using four composite scores (Revilla 2010: 151). However, since allocation to modes was not random within this design, it is not certain that the mode of administration was indeed responsible for these differences, rather than characteristics of the respondents, which may be related both to their propensity to give less reliable and valid answers and to their likelihood to choose a given mode of data collection.

Self-completion questionnaires

Self-completion questionnaires (SAQ), including computer assisted self-completion questionnaires (CSAQ), have been demonstrated to differ from face-to-face interviews in the degree to which they elicit socially undesirable responses. Respondents are more likely to report socially undesirable opinions and behaviours when filling in a self-completion questionnaire than when responding to a face-to-face interview. In several cases, where researchers have compared estimates derived from surveys to estimates derived from official statistics or objective tests, it could be shown that self-completion questionnaires in general lead to more accurate estimates than face-to-face interviews (Tourangeau and Smith 1996: 275).

Although the smaller social desirability bias of SAQs is well established, the sizes of the differences to face-to-face interviews vary with the particular survey question asked. Moreover, while the two methods lead to different results on *some* attitude items with empirically established social desirability connotations, there appears to be no differential measurement effect on others (Heerwegh 2009). For surveys focusing on attitudes, in particular, it thus becomes difficult to predict which questions would be affected by mode effects.

Although it is difficult to establish the accuracy of an estimate derived from an attitude question (for unlike with some behavioural items, it is difficult to prove that a given response to an attitude question is “wrong” or “right”), it is often reasonable to assume that the greater privacy that SAQs afford the respondent relative to face-to-face interviews leads to more accurate reports of proscribed attitudes. Yet as I have said before, improved accuracy is not always an advantage, because it introduces a bias into comparisons with countries that do not employ SAQs, or with earlier data collection waves using other modes.

The absence of an interviewer in SAQs has also raised concerns about the quality of the data collected. Skilled interviewers can help respondents understand questions they find difficult – thereby reducing task difficulty – and can raise respondent motivation through interaction and encouraging feedback. Moreover, respondents who answer questionnaires via the web may engage in a variety of other activities simultaneously (such as visiting other websites, answering e-mails, ...), and may therefore be less attentive to the task. It is thus likely that web questionnaires may be more prone than face-to-face interviews to encourage respondents to satisfice, and to leave items unanswered.³

With the exception of research into primacy and recency effects, there are few studies that have compared self-completion and face-to-face modes in terms of data quality (Roberts 2007: 14). Moreover, the effect of mode on satisficing and item non-response is likely to vary by length of questionnaire (Roberts 2007). One study carried out under the Mixed Mode Methodology Programme compared data from face-to-face interviews and web questionnaires in a relatively long questionnaire (180 to 235 survey questions) in two random samples of university students (Heerwegh 2009; Heerwegh and Loosveldt 2008: 836). The analysis showed that web respondents were significantly more likely than face-to-face interviewees to answer “don’t know”, to fail to respond to individual items, and to use the middle category of response scales (an indicator of non-differentiation). While these results constitute evidence for the hypothesis that SAQs might be more likely to induce satisficing than face-to-face interviews, the differences between the modes had only small effects on the substantive distributions of answers.

WCMM designs

The measurement effects described above have the potential to create particular problems when more than one mode is used within a single country. This is because in any practicable WCMM design, it is very unlikely that the allocation of respondents to different modes will be governed by a random process. Rather, we expect that the selected mode of data collection will be related to respondent characteristics. This would mean that comparisons of subgroups of a national

population may potentially be subject to measurement bias, if different subgroups have different probabilities to select into the available modes, and if the dependent variable is subject to measurement effects due to mode. Such within-country effects would, of course, also have implications for cross-country comparisons involving population subgroups.

The implications of WCMM designs for cross-country comparisons are also potentially severe. Even where two countries combine the same modes (say, face-to-face and web), the designs would likely be of uncertain equivalence, because a number of conditions are likely to vary by country: sampling frames used for different modes might be different, and so might mode penetration (e.g. internet access) and sampling practicalities. This means that *the processes of selection by which respondents end up responding in one mode or another* are likely to differ from country to country. Quite simply, countries will inevitably differ in the proportions of sample members who take up one mode or another. Once again, we would face the possibility of biased comparisons between countries.

Another consideration concerns the impact of WCMM designs on the analysis of change. The introduction of WCMM in a country would not only have implications for comparisons of WCMM waves of data collection with previous waves conducted as single-mode face-to-face surveys. It would also endanger the equivalence of all subsequent WCMM-waves with one another. This is because it is unlikely that two subsequent waves of WCMM data collection would be identical in terms of the process of selection into different modes, and the proportion of respondents taking up the various modes.⁴ A country introducing WCMM, then, would in fact become an ATMM country with a (potentially) *different* design in each wave.

Unknown implications of mixed mode designs

Despite the considerable body of evidence on measurement effects that I have reviewed so far, our ignorance concerning the effects of mixing modes of data collection on measurement precision still far outweighs our knowledge. Although some measurement effects have been shown to have affected data in some countries, with some variables, and at certain points in time, there is no guarantee that these effects will be the same or similar in other countries, or stable over time, or that we can extrapolate from the evidence concerning given variables to the likely effects on other variables. Therefore, the introduction, without further testing, of mixed mode designs (whether ACMM only, or WCMM) into a cross-national survey would most likely mean that data analysts who wanted to take into account mode effects would, to a certain extent, have to rely on untested assumptions about the existence, size and (geographical and temporal) invariance of mode effects.

Coverage

Modes of data collection differ in the extent to which they approximate full coverage of national populations. Face-to-face interviews, where interviewers visit respondents in their homes and can also record information about non-eligible addresses and non-respondents, are usually considered to offer the best coverage, and the best chance of collecting paradata. Postal questionnaires may also provide good coverage, although they do not offer the chance to collect paradata, and suffer from low response rates relative to face-to-face interviews.

Telephone interviews

Traditionally, telephone surveys have been carried out on fixed-line telephones. Households were sampled either by selecting telephone numbers from a list, or by one of several methods of random digit dialling (RDD) (see Tourangeau 2004: 777–781). In both cases, it is possible to achieve a probability sampling design. However, the rise of mobile phone technology has meant that an increasing number of people abandon fixed-line telephones completely, and thus are not covered by these existing sampling frames. For example, data from the European Social Survey (Round 3) indicate that very few participating countries have sufficient “penetration” of fixed-line telephones to carry out a single mode telephone survey relying on fixed-line telephones only.

Of the countries participating in ESS Round 3, only Switzerland had a fixed-line telephone coverage of more than 95% of the population (Roberts, Eva and Widdop 2008). Only five other countries had at least 90% coverage. Moreover, the trend is in the direction of decreasing fixed-line telephone access. A comparison of data from ESS Rounds 3 and 4 indicates that fixed-line telephone coverage declined in most countries between 2006 and 2008. In 2008, only three countries had an estimated fixed-line coverage of more than 90% (Sweden, Germany, and France).⁵

On balance, it is fair to say that most countries could not operate a single-mode telephone survey relying on fixed-line telephones without departing drastically from the principle of “full coverage of the population” that should guide the sampling strategy of cross-national surveys. However, the coverage problem on its own does not necessarily prevent telephone interviews from being part of a WCMM data collection design.

Web questionnaires

There is currently no country in the world where internet coverage is sufficiently high to conduct an internet-only survey and satisfy the requirement of full coverage of the national population (Lozar and Vehovar 2008: 264).⁶ A single-mode web-based

survey is therefore not yet realistic. Again, however, this does not disqualify web questionnaires of playing a part in a WCMM design.

However, recently the practice of web panels for regular survey participation has become popular (Duffy et al. 2005: 615; Bandilla, Bosnjak and Altdorfer 2003: 235). Individuals are selected randomly from the population, and invited to join a panel of respondents that are then regularly asked to complete online questionnaires. In some cases, as with the Dutch LISS panel (<http://www.lissdata.nl/lissdata/>), volunteers without internet access or computers are provided with the necessary equipment. This way, researchers hope to establish a cost-effective way of reaching a sample of respondents that may represent the population as well as a freshly drawn random sample. However, two major drawbacks cast doubt on the suitability of web panels for cross-national surveys: first, participation of panel members in surveys may drop over time, so that the sample as a whole would be in danger of becoming less representative of the population as time goes on. Second, regular participation in surveys may itself have an effect on the respondents' answers, as survey habituation may have psychological consequences, such as greater sophistication in reading and answering survey questions, or survey boredom.

Non-response

In general, survey non-response may be attributable to three types of causes: non-contact; refusal to respond; and inability to respond. Evidence from the USA suggests that all three causes have a part to play in the explanation of falling response rates (Tourangeau 2004: 782).⁷

Telephone interviews

As a single mode, telephone surveys of national populations are generally regarded as prone to lower response rates than face-to-face surveys (Groves and Lyberg 1988: 203). However, response rates may well vary between countries, partly due to differing "survey taking climates" (Loosveldt and Storms 2008: 74) that make different modes of contact more or less familiar and acceptable to a country's population. Thus, in a consultation with ESS field directors, telephone was estimated to produce higher response rates than face-to-face in seven out of twenty-three countries (Roberts, Eva and Widdop 2008: 41).⁸

Yet the results of an experiment conducted as part of the ESS Mixed Mode Methodology Programme suggest that response rates would suffer if the ESS were conducted by telephone. In four countries (Germany, Hungary, Poland, and the two French regions of Switzerland), a telephone survey achieved significantly lower response rates than the ESS 3rd round mainstage survey conducted at the same time

(Roberts, Eva, Lynn and Johnson 2009). The differences in response rates between the full ESS survey interview conducted by telephone and the same interview conducted by face-to-face ranged from 8 percentage points in Switzerland to 48 percentage points in Hungary. It is possible that some telephone surveys can achieve relatively high response rates because they are relatively short, whereas in the case of the ESS, which takes around an hour, interviewers find it difficult to persuade sample members to participate. Yet attempts to increase the telephone response rate by offering shorter interviews (45 mins, or two stages of 30 mins each) still did not bring the response rates up to the level of the face-to-face interviews.

Self-administered questionnaires

Single-mode postal and web surveys also frequently suffer from high rates of non-response, and there is usually a high degree of uncertainty about the causes of non-response (for if there is no response, survey researchers usually have no feasible means of establishing whether an e-mail or a postal questionnaire ever reached its intended recipient) (Vehovar, Batagelj, Lozar Manfreda and Zaletel 2002: 229). However, in a WCMM design the first contact with respondents would be likely to be established by methods other than e-mail, so this weakness of SAQ modes would not be an issue.

WCMM designs

When methodologists write about mixed mode data collection, they often express the hope that a mixed mode approach will help to mitigate falling response rates. In particular, following up non-contacts in a different mode is sometimes thought to decrease the non-contact rate, and offering sample members a choice of mode carries the hope of reducing refusal rates, by offering people the chance to answer in the mode they personally prefer.

However, it is doubtful whether mixed mode data collection by itself would contribute to enhancing response rates (or at least to halting the declining trend). Evidence suggest that face-to-face data collection is the single mode likely to achieve the highest response rates. Mixed modes *can* enhance response rates relative to modes *other* than face-to-face interviews; for example, when non-respondents to an initial mail are followed up in an interviewer-administered mode.⁹ There is, however, no evidence that mixed mode data collection will enhance response rates relative to single mode *face-to-face* surveys. In particular, giving respondents a choice of mode does not appear to raise response rates, and if anything, sometimes even decreases it (Dillman, Smyth and Christian 2009: 304 ff).¹⁰

Neither does evidence from the experiments carried out under the ESS's own Mixed Mode Methodology Programme support the hope that mixed mode

designs can raise response rates. In a mixed mode experiment carried out in the Netherlands parallel to Round 4 of the ESS, response rates in two experimental groups implementing different mixed mode designs were around 8 percentage points *lower* than the response rate in the regular Round 4 face-to-face survey (Eva et al. 2010).¹¹ It is possible that the implementation of the WCMM design in this experiment may be improved upon to raise response rates.¹² Yet it seems that a WCMM design could at best hope to match the response rate of a face-to-face survey, rather than exceed it.

Practical complexity

Mixed mode designs are more complex than single-mode designs from a practical perspective. Special specifications, questionnaires and fieldwork documents have to be produced for each mode (even if a particular mode is only applied in one ESS country), data from diverging question formats have to be made compatible across modes, and special documents advising data analysts of the possibility of measurement effects due to mode have to be prepared and published along with the data.

The real complexity starts, however, when we consider WCMM designs. The most likely WCMM designs will be those that involve non-respondents to a contact attempt in one mode who are then followed up in a different mode. This means that fieldwork will have to be carefully organized so that respondents are “switched” from one mode to another, as well as, possibly, back again in the course of the fieldwork period. As fieldwork gets more complex, the need for additional fieldwork monitoring by national coordinating teams arises, further increasing the workload of a WCMM design relative to a single-mode design.

An example from the ESS Mixed Mode Methodology Programme illustrates how difficult it can be to implement a rigorous WCMM design. In the WCMM experiment conducted in the Netherlands parallel to ESS Round 4 (already referred to above), respondents were offered (either concurrently or sequentially) a choice of three modes: a web self-completion questionnaire, telephone interviews, or face-to-face interviews. Where telephone numbers were available, sample members were contacted by telephone first, and depending on their expressed preference, completed a telephone interview on the spot or arranged a telephone interview at a later time, arranged a visit by a personal interviewer, or arranged to receive the link to an online questionnaire. Of those who chose web, however, less than two thirds completed the questionnaire after receiving up to ten telephone reminders. Ideally, these respondents should have been followed up by a visit from a face-to-face interviewer, in order to maximize the response rate. But they were not, and the fieldwork process was not monitored closely enough to prevent this oversight.

While such problems are not impossible to overcome, it is clear that doing so requires meticulous planning as well as additional resources during fieldwork.

EVALUATING MIXED MODE DATA COLLECTION

The investigation of mode effects has shown that a cross-national survey that employs a mixed mode design (whether ATMM, ACMM, or WCMM) is unlikely to attain the same level of equivalence that a single-mode design would. To make matters more complicated, mode of measurement may affect different variables in different ways – some not at all, some weakly, others more strongly – and for a given survey, the effects on all types of variables will be difficult to predict. It is of course useful to have quantified the social desirability bias associated with different modes of data collection on variables such as drug use and reports of sexual encounters in US American samples (Tourangeau and Smith 1996: 275). But this does not tell us the effect mixed mode designs would exert on other variables, and in samples from other populations. Even in cases where we have strong theoretical reasons for expecting certain kinds of effect (say, in the case of sensitive questions, mode-specific social desirability biases), we usually have little indication about the effect size. Thus, the internal validity of cross-national comparisons would be in doubt.

Surveys differ in the types of questions they ask, and the topics they focus on. It is therefore difficult to deduce conclusions from the scientific literature that apply to all surveys that may consider employing mixed mode data collection. The best approach, from a methodological point of view, to introducing mixed mode data collection would be to incorporate a methodological experiment into the survey, whereby any country whose data collection design diverges from the dominant mode within the cross-national survey would be required to run a parallel survey in the dominant mode (with a smaller sample). Parallel designs have been identified as good practice in survey transitions – when an important design element of a repeat cross-sectional survey is to be changed (van den Brakel, Jan A., Smith and Compton 2008, p. 123). Such methodological experiments would allow researchers to gauge the overall effect of the mixed mode design.

CONCLUSION

We have seen that different survey modes may lead to differences in measurements. Therefore, an international survey that allows different participating countries to collect data in different modes risks compromising the very basis of comparative measurement: namely the equivalence of the measurement instrument. Moreover, there is little evidence that within-country mixed mode (WCMM) surveys can improve survey features so as to improve precision of measurement, since hopes

in the raising of response rates (relative to face-to-face interviews) do not seem to be confirmed by the available evidence. Therefore, from the point of view of the data analyst, mixed mode surveys introduce a considerable disadvantage, without having a measurement advantage.

On the other hand, it is true that all cross-national surveys need to have a pragmatic approach to equivalence, since almost all surveys are subject to a great number of potential measurement effects (due to, for example, imperfect population coverage, country-specific non-response error, questionnaire translation errors, country-specific interviewer errors, and so forth) that will probably never be rooted out completely, and whose effects on the data will often remain unknown, as it would take too many resources to methodologically control every possible threat to equivalent measurement.

Creative solutions to some of the methodological problems posed by mixed mode data collection could be found. For example, if a survey was to mix face-to-face interviews and self-completion modes, it might be possible to reduce some of the mode differences by introducing a self-completion section into the face-to-face interview, which would contain sensitive questions and items known to be affected by social desirability bias. In accordance with the definitions of Section 2, this would in effect be a combination of mixed mode and multiple mode data collection. The method does not, however, deal with lack of diachronic equivalence introduced if a time series survey switches modes from one wave to another.

Pressures of fieldwork costs and dwindling response rates may make mixed mode designs appear attractive for cross-national surveys, yet there is a danger that such designs be implemented without proper methodological reflection and research. This may introduce uncontrolled and unknown biases into the data, which may harm the internal validity of cross-country comparisons. At a time when large surveys strive to make their data available to a wide range of users, and when large-scale survey data are stored in databases that may be analysed not just concurrently, but longitudinally for years and decades to come, we can hardly overemphasize the risk associated with the unthinking introduction of mixed mode designs.

This is especially pertinent when we consider that large-scale survey data are going to be used by a wide variety of users. While statistically and methodologically sophisticated users may exercise caution when analysing data from mixed mode designs, and consult methodological evidence to gauge the likelihood and extent of potential biases, other types of users may well take the data at face value. In a climate where it is often difficult to make non-scientists understand even relatively simple notions such as survey weights or confidence intervals, it would be rather over-optimistic on the part of the data provider to assume that all users would know to exercise caution when interpreting mixed mode data.

In any case, it is not necessarily clear automatically what ‘exercising caution’ is supposed to mean. Without knowledge of the biases that mixed mode designs would introduce, corrections are impossible to make, and apart from including in any publication a generic ‘health warning’ about the doubtful validity of country comparisons, there is nothing the researcher can do. The introduction of mixed mode designs without methodological knowledge specific to the demands of the particular survey, and without detailed advice for data users, is likely to result in the provision of flawed data that may not only harm the validity of results gained from them, but in the long run may tarnish the reputations of cross-national surveys themselves.

NOTES

- 1 For a summary of the ESS Mixed Mode Methodology Programme see: <http://www.europeansocialsurvey.org>.
- 2 The National Child Development study (NCDS) follows all people born in a certain week in March 1958; the British Cohort Study (BCS70) follows all people born in a certain week in April 1970 (cf. www.cls.ioe.ac.uk/).
- 3 Roberts (2007: 14) argues that self-completion respondents may feel less time pressure than interview respondents, and that they might therefore be encouraged to answer questions more carefully, leading to less satisficing (or possibly offsetting some of the other characteristics of web surveys). Yet in a survey experiment by Heerwegh and Loosveldt (2008: 841), web respondents actually completed a survey much faster than respondents to a face-to-face interview. In their study, web-respondents took an average of 32 minutes to answer between 180 and 235 questions, while face-to-face interviews took an average of 48 minutes.
- 4 For example, let’s assume that a country adopts a two-mode design, combining face-to-face interviews (CAPI) and web questionnaires (WSAQ). Let’s further assume that in the first ESS wave adopting this design, 30% of respondents fill in a WSAQ, and 70% do a CAPI. If these proportions change in the next wave (say, to 60% taking the WSAQ, and 40% doing a CAPI), then any variables affected by measurement effects due to mode would not have been measured equivalently across waves, even though both waves seem to have, on first sight, been conducted with the same data collection design. Incidentally, since cost reduction is the chief motivator for mixed mode designs, a rising proportion of WSAQ respondents and a decreasing proportion of CAPI respondents would be in both the funders’ and the survey agency’s interest, and would also be likely as younger, more internet-savvy generations replace older ones. Changing mode proportions are thus not just a possibility, but are to be expected.
- 5 According to ESS Round 4, the Swiss fixed-line coverage has declined dramatically. In 2008, only 89% of Swiss households were estimated to have a fixed-line telephone. This apparent decline may in fact be an artefact of a change in sampling frame in Switzerland between ESS rounds 3 and 4. Round 3 relied on a telephone register for its sampling, and may therefore have been biased towards households with a fixed-line telephone.
- 6 However, internet coverage is increasing fast, and near-full coverage may or may not soon be realistic in some countries. The only studies that manage to rely on the internet for all data collection *and* can hope to achieve a probability sample from the population

are studies that provide hardware, software and internet access for panel respondents who would not otherwise have the necessary equipment to participate (Lozar Manfreda, Bosnjak, Berzelak, Haas, et al. 2008: 265). Due to the high cost of realizing such a design, it is extremely unlikely that it would ever be feasible within the ESS.

- 7 De Leeuw and de Heer (2002) provide evidence that both non-contact and refusal increased between the 1970s and the 1990s across different types of surveys in sixteen developed countries, including the USA, Canada, Australia, and thirteen European countries. The authors do not provide evidence on the rates of inability to participate.
- 8 These seven countries were: Denmark, Finland, France, Iceland, Norway, Sweden, and Switzerland.
- 9 This is practiced in the American Community Survey, which was referred to above.
- 10 The evidence Dillman et al. cite largely stems from comparisons of single mode postal surveys with mixed postal and web surveys.
- 11 Response Rates: ESS Round 4 Netherlands (face-to-face): 52.0%; concurrent mixed mode design (offering respondents a choice of three modes: face-to-face, telephone, and web): 44.0%; sequential mixed mode design (offering respondents web first, and offer those who refuse telephone and face-to-face, in that order): 43.4%.
- 12 A potential improvement might be achieved through better follow-up of respondents who agreed to fill in an online questionnaire, but do not do so even after several reminders.

REFERENCES

- Bandilla, Wolfgang, Michael Bosnjak and Patrick Altdorfer. 2003. 'Survey Administration Effects? A Comparison of Web-Based and Traditional Written Self-Administered Surveys using the ISSP Environment Module.' *Social Science Computer Review* 21 (2): 235–243.
- de Leeuw, Edith D. 2005. 'To Mix Or Not to Mix Data Collection Mode in Surveys.' *Journal of Official Statistics* 21 (2): 233–255.
- de Leeuw, Edith D. and Wim de Heer. 2002. 'Trends in Household Survey Nonresponse: A Longitudinal and International Comparison.' In: *Survey Nonresponse*, edited by R.M. Groves, D.A. Dillman, J.L. Eltinge and R.J.A. Little. New York: Wiley.
- de Leeuw, Edith D., Don A. Dillman and Joop J. Hox. 2008. 'Mixed Mode Surveys: When and Why.' Pp. 299–316. In: *International Handbook of Survey Methodology*, edited by E.D. de Leeuw, J.J. Hox and D.A. Dillman. Boca Raton: CRC Press.
- Dillman, Don A., Jolene D. Smyth and Leah Melani Christian. 2009. *Internet, Mail, and Mixed Mode Surveys: The Tailored Design Method*. Hoboken, NJ: Wiley.
- Duffy, Bobby, et al. 2005. 'Comparing Data from Online and Face-to-Face Surveys.' *International Journal of Market Research* 47 (6): 615–639.
- Eva, Gillian, Geert Loosveldt, Peter Lynn, Peter Martin, Melanie Revilla, Willem Saris and Jorre Vannieuwenhuyze (2010) 'ESS Prep WP6 – Mixed mode experiment. Final mode report'. Unpublished research report. Centre for Comparative Social Surveys, City University, London.
- Fitzgerald, Rory and Roger Jowell. 2011. 'Measurement Equivalence in Comparative Surveys: The European Social Survey (ESS) – from Design to Implementation and Beyond.' CCSS Working Paper No. 3. Centre for Comparative Social Surveys, City University, London. Available: www.europeansocialsurvey.org [Access 6/12/2011].

- Groves, Robert M., et al. 2004. *Survey Methodology*. Hoboken, NJ: Wiley.
- Groves, Robert M. and Lars E. Lyberg. 1988. 'An Overview of Nonresponse Issues in Telephone Surveys.' Pp. 191–211. In: *Telephone Survey Methodology*, edited by R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls II and J. Waksberg. New York: Wiley.
- Heerwegh, Dirk. 2009. 'Mode differences between face-to-face and web surveys: an experimental investigation of data quality and social desirability effects'. *International Journal of Public Opinion Research* 21: 111–121.
- Heerwegh, Dirk and Geert Loosveldt. 2008. 'Face-to-Face Versus Web Surveying in a High-Internet Coverage Population. Differences in Response Quality.' *Public Opinion Quarterly* 72 (5): 836–846.
- Holbrook, Allyson L., Melanie C. Green and Jon A. Krosnick. 2003. 'Telephone Versus Face-to-Face Interviewing of National Probability Samples with Long Questionnaires: Comparisons of Respondent Satisficing and Social Desirability Bias.' *Public Opinion Quarterly* 67 (1): 79–125.
- Jäckle, Annette, Caroline Roberts and Peter Lynn. 2010. 'Assessing the Effect of Data Collection Mode on Measurement.' *International Statistical Review* 78 (1): 3–20.
- . 2006. 'Telephone Versus Face-to-Face Interviewing: Mode Effects on Data Quality and Likely Causes. Report on Phase II of the ESS-Gallup Mixed Mode Methodology Project.' Unpublished report. Centre for Comparative Social Surveys, City University, London.
- Jowell, Roger. 1998. 'How Comparative is Comparative Research?' *American Behavioral Scientist* 42 (2): 168–177.
- Jowell, Roger, Max Kaase, Rory Fitzgerald and Gillian Eva. 2007. 'The European Social Survey as a Measurement Model.' Pp. 1–31. In: *Measuring Attitudes Cross-Nationally. Lessons from the European Social Survey*, edited by R. Jowell, C. Roberts, R. Fitzgerald and G. Eva. Los Angeles: Sage.
- Loosveldt, Geert and Vicky Storms. 2008. 'Measuring Public Opinion about Surveys.' *International Journal of Public Opinion Research* 20 (1): 74–89.
- Lozar Manfreda, Katja, et al. 2008. 'Web Surveys Versus Other Survey Modes. A Meta-Analysis Comparing Response Rates.' *International Journal of Market Research* 50 (1): 79–104.
- Lozar Manfreda, Katja and Vasja Vehovar. 2008. 'Internet Surveys.' Pp. 264–284. In: *International Handbook of Survey Methodology*, edited by E.D. De Leeuw, J.J. Hox and D.A. Dillman. Boca Raton: CRC Press.
- Pew Global Attitudes Project. 2009. 'Pew Global Attitudes Project. Pew Research Centre. Website: <http://www.pewglobal.org/> [Access 6/12/2011].
- Revilla, Melanie. 2010. 'Quality in Unimode and Mixed-Mode Designs: A Multi-Trait Multimethod Approach.' *Survey Research Methods* 4 (3): 151–164.
- Roberts, Caroline. 2007. 'Mixing Modes of Data Collection in Surveys: A Methodological Review.' ESRC National Centre for Research Methods. NCRM Methods Review Papers, NCRM/008. Available: <http://eprints.ncrm.ac.uk/418/> [Access 6/12/2011].
- Roberts, Caroline, Gillian Eva, Peter Lynn, and Jerry Johnson. 2009. 'Measuring the effect of interview length on response propensity and response quality in a telephone

- survey. Final report of the ESS CATI experiment'. Unpublished report, Centre for Comparative Social Surveys, City University, London.
- Roberts, Caroline, Gillian Eva and Sally Widdop. 2008. 'Assessing the Demand and Capacity for Mixing Modes of Data Collection on the European Social Survey: Final Report on the Mapping Exercise.' Unpublished report. Centre for Comparative Social Surveys, City University, London.
- Tourangeau, Roger. 2004. 'Survey Research and Societal Change.' *Annual Review of Psychology* 55: 775–801.
- Tourangeau, Roger and Tom W. Smith. 1996. 'Asking Sensitive Questions. the Impact of Data Collection Mode, Question Format and Question Context.' *Public Opinion Quarterly* 60: 275–304.
- US Census Bureau. 2009. 'Design and Methodology: American Community Survey.' Available: http://www.census.gov/acs/www/Downloads/survey_methodology/acs_design_methodology.pdf [Access: 6/12/2011].
- van den Brakel, Jan A., Paul A. Smith and Simon Compton. 2008. 'Quality Procedures for Survey Transitions – Experiments, Time Series and Discontinuities.' *Survey Research Methods* 2 (3): 123–141.
- Vehovar, Vasja, Zenel Batagelj, Katja Lozar Manfreda and Metka Zaletel. 2002. 'Nonresponse in Web Surveys.' Pp. 229–242. In: *Survey Nonresponse*, edited by R.M. Groves, D.A. Dillman, J.L. Eltinge and R.J.A. Little. New York: Wiley.
- World Values Survey. 2009. *Technical Specifications of the Values Surveys World Values Survey*. Available: www.wvsevsdb.com/wvs/WVSTechnical.jsp [Access: 6/12/2011].

Peter Martin teaches research methods and statistics at the Anna Freud Centre (London, UK) and University College London. Previously he was a research fellow at City University, London, and a member of the Central Coordinating Team of the European Social Survey. He is also interested in the study of racism, the psychology of moral action, mixed methods research, and sequence analysis. His recent publications include: Martin, P. & Wiggins, R. D. (2011) 'Optimal matching analysis'. In: Williams, M. & Vogt, W. P. (eds) *The Sage handbook of innovation in social research methods*. London: Sage (pp. 385–408). Martin, P.; Schoon, I. & Ross, A. (2008) 'Beyond transitions. Applying optimal matching analysis to life course research'. *International Journal of Social Research Methodology: Theory and Practice* 11 (3): 179–199.

The Anna Freud Centre, 12 Maresfield Gardens, London NW3 5SU
 Email: peter.martin@annafreud.org