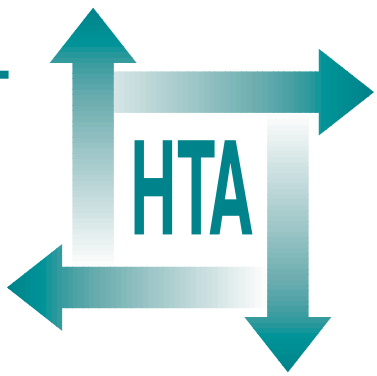
Health Technology Assessment 2001; Vol. 5: No. 31

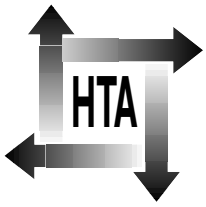
Methodology

Design and use of questionnaires: a review of best practice applicable to surveys of health service staff and patients

E McColl
A Jacoby
L Thomas
J Soutter
C Bamford
N Steen
R Thomas
E Harvey
A Garratt
J Bond

Health Technology Assessment
NHS R&D HTA Programme





How to obtain copies of this and other HTA Programme reports.
An electronic version of this publication, in Adobe Acrobat format, is available for downloading free of charge for personal use from the HTA website (http://www.ncchta.org). A fully searchable CD-ROM is also available (see below).

Printed copies of HTA monographs cost £20 each (post and packing free in the UK) to both public **and** private sector purchasers from our Despatch Agents, York Publishing Services.

Non-UK purchasers will have to pay a small fee for post and packing. For European countries the cost is £2 per monograph and for the rest of the world £3 per monograph.

You can order HTA monographs from our Despatch Agents, York Publishing Services by:

– fax (with **credit card** or **official purchase order**)
– post (with **credit card** or **official purchase order** or **cheque**)
– phone during office hours (**credit card** only).

Additionally the HTA website allows you **either** to pay securely by credit card **or** to print out your order and then post or fax it.

Contact details are as follows:

York Publishing Services
PO Box 642
YORK YO31 7WX
UK

Email: ncchta@yps-publishing.co.uk
Tel: 0870 1616662
Fax: 0870 1616663
Fax from outside the UK: +44 1904 430868

NHS and public libraries can subscribe at a very reduced cost of £100 for each volume (normally comprising 30–40 titles). The commercial subscription rate is £300 per volume. Please contact York Publishing Services at the address above. Subscriptions can only be purchased for the current or forthcoming volume.

Payment methods

Paying by cheque
If you pay by cheque, the cheque must be in **pounds sterling,** made payable to *York Publishing Distribution* and drawn on a bank with a UK address.

Paying by credit card
The following cards are accepted by phone, fax, post or via the website ordering pages: Delta, Eurocard, Mastercard, Solo, Switch and Visa. We advise against sending credit card details in a plain email.

Paying by official purchase order
You can post or fax these, but they must be from public bodies (i.e. NHS or universities) within the UK. We cannot at present accept purchase orders from commercial companies or from outside the UK.

How do I get a copy of *HTA on CD*?

Please use the form on the HTA website (www.ncchta.org/htacd.htm). Or contact York Publishing Services (see contact details above) by email, post, fax or phone. *HTA on CD* is currently free of charge worldwide.

The website also provides information about the HTA Programme and lists the membership of the various committees.

Design and use of questionnaires: a review of best practice applicable to surveys of health service staff and patients

E McColl[1]*
A Jacoby[1]
L Thomas[1]
J Soutter[1]
C Bamford[1]
N Steen[1]
R Thomas[2]
E Harvey[3]
A Garratt[3]
J Bond[1]

[1] Centre for Health Services Research, University of Newcastle, UK
[2] National Centre for Social Research, London, UK
[3] Department of Health Sciences, University of York, UK

* Corresponding author

Competing interests: none declared

Published December 2001

This report should be referenced as follows:

McColl E, Jacoby A, Thomas L, Soutter J, Bamford C, Steen N, *et al*. Design and use of questionnaires: a review of best practice applicable to surveys of health service staff and patients. *Health Technol Assess* 2001;**5**(31).

Health Technology Assessment is indexed in *Index Medicus*/MEDLINE and *Excerpta Medica*/EMBASE. Copies of the Executive Summaries are available from the NCCHTA website (see opposite).

NHS R&D HTA Programme

The NHS R&D Health Technology Assessment (HTA) Programme was set up in 1993 to ensure that high-quality research information on the costs, effectiveness and broader impact of health technologies is produced in the most efficient way for those who use, manage and provide care in the NHS.

Initially, six HTA panels (pharmaceuticals, acute sector, primary and community care, diagnostics and imaging, population screening, methodology) helped to set the research priorities for the HTA Programme. However, during the past few years there have been a number of changes in and around NHS R&D, such as the establishment of the National Institute for Clinical Excellence (NICE) and the creation of three new research programmes: Service Delivery and Organisation (SDO); New and Emerging Applications of Technology (NEAT); and the Methodology Programme.

Although the National Coordinating Centre for Health Technology Assessment (NCCHTA) commissions research on behalf of the Methodology Programme, it is the Methodology Group that now considers and advises the Methodology Programme Director on the best research projects to pursue.

The research reported in this monograph was funded as project number 93/49/02.

The views expressed in this publication are those of the authors and not necessarily those of the Methodology Programme, HTA Programme or the Department of Health. The editors wish to emphasise that funding and publication of this research by the NHS should not be taken as implicit support for any recommendations made by the authors.

Criteria for inclusion in the HTA monograph series
Reports are published in the HTA monograph series if (1) they have resulted from work commissioned for the HTA Programme, and (2) they are of a sufficiently high scientific quality as assessed by the referees and editors.

Reviews in *Health Technology Assessment* are termed 'systematic' when the account of the search, appraisal and synthesis methods (to minimise biases and random errors) would, in theory, permit the replication of the review by others.

The editors and publisher have tried to ensure the accuracy of this report but do not accept liability for damages or losses arising from material published in this report. They would like to thank the referees for their constructive comments on the draft document.

ISSN 1366-5278

Published by Core Research, Alton, on behalf of the NCCHTA.
Printed on acid-free paper in the UK by The Basingstoke Press, Basingstoke.

M

Contents

List of abbreviations

admin.	administration*
ANOVA	analysis of variance*
CAPI	computer-assisted personal interviewing
CASI	computer-assisted self-administration
CATI	computer-assisted telephone interviewing
CI	confidence interval
DISKQ	diskettes programmed with a questionnaire
DK	"don't know" response*
FPC	Family Practitioner Committee
MCG	Morality–Conscience–Guilt (scale)*
MMDQ	Moos Menstrual Distress Questionnaire*
ns	not significant*
OCR	optical character recognition
OMR	optical mark reading
PG	postgraduate*
PPI	printed postage impressions
RCT	randomised controlled trial*
RR	relative risk
SD	standard deviation*
SES	socio-economic status*
SF-36	Medical Outcomes Study Short Form-36 (questionnaire)
UG	undergraduate*

* Used only in tables

Executive summary

Introduction

Questionnaires are often used to collect primary quantitative data from patients and healthcare professionals. The aim is to gather valid, reliable, unbiased and discriminatory data from a representative sample of respondents. However, the information yielded is subject to error and bias from a range of sources. Close attention to issues of questionnaire design and survey administration can reduce these errors.

Objectives

A selective, narrative literature review was conducted to identify current best practice with respect to the design and conduct of questionnaire surveys, including theories of respondent behaviour, "expert opinion" and high-quality evidence from experimental studies. The principal foci were:

- modes of survey administration (various forms of interviewer administration and self-completion)
- question wording, choice of response formats, and question sequencing
- questionnaire formatting and other aspects of presentation
- techniques for enhancing response rates, with particular emphasis on postal surveys.

Methods of the review and implications for readers

The starting point for this review was "expert opinion", encapsulated in key textbooks on the design and conduct of surveys. High-grade evidence was then sought from experimental and quasi-experimental studies to support or refute the experts' recommendations. In addition, information was sought on the theoretical underpinnings of survey response. A deliberate and considered decision was made to include studies from disciplines other than health because it was envisaged that theories of respondent behaviour, as well as methodological messages, are unlikely to be discipline specific. The PsycLIT electronic database was therefore used in addition to MEDLINE, but the search was confined to articles published in the English language between 1975 and 1996. It is acknowledged that confining the search to two databases only is likely to have led to bias in favour of articles published in the major American and British journals indexed on those databases, and to exclusion of the "grey" literature.

Owing to human error, references identified through MEDLINE for the period 1987–1992, and those identified through PsycLIT for 1979, 1991 and 1993–1996, were excluded from consideration (appendix 5). Although it is acknowledged that these omissions mean that this review cannot be considered to be systematic, the authors believe that their conclusions would not have been materially altered by the incorporation of articles identified through the two key databases for the years in question. In contrast to clinical research, where the accretion of knowledge tends to be incremental, with new studies seeking to replicate or refute the findings of those that have gone before, research into questionnaire construction and survey administration appears to be haphazard, often with little reference to previous studies. Examination of the literature provided little sense of concerted efforts to generalise findings from one study to other settings, populations or modes of administration.

Explicit inclusion, exclusion and quality criteria were applied in a two-stage process of sifting and then synthesising findings from identified studies. However, because of the heterogeneity of studies, no attempts at meta-analysis were made. To facilitate comparisons between studies, findings are presented as relative risks with associated 95% confidence intervals (for differences in percentages), or as differences in means with associated 95% confidence intervals (for continuous data). In setting out the findings, a distinction has also been made between studies on health-related topics and those from other fields. A quality score is included for each identified study.

In defining the scope of this review, an explicit decision was made to exclude certain aspects of the survey process, most notably sampling and pilot testing. These features of survey methodology do not lend themselves readily to experimental

investigation and they are likely to be highly study and topic specific. Indeed, in the context of health technology assessment, definitions of sample inclusion and exclusion criteria and sample size calculations are likely to be predicated on the design of the parent trial. However, in recognition of the importance of these aspects of the survey process, appendix 1 provides brief guidance on key topics omitted from the formal review. This appendix is based primarily on the collective experience of the authors, with limited references to key texts and articles on the chosen topics.

Results

Mode of administration

The two principal modes of administration are self-completion and interviewer administration. Evidence from identified studies provided no consistent picture of the superiority of any one mode in terms of the quantity or quality of the response, or the resources required.

Question wording and sequencing

Evidence from identified studies supported the notion that question wording and framing, including the choice and order of response categories, can have an important impact on the nature and quality of responses.

The conventional wisdom with respect to question ordering is that general questions should precede specific questions; evidence from a number of primary studies supported this assertion.

Questionnaire appearance

Through careful attention to the design and layout of questionnaires, the risk of errors in posing and interpreting questions and in recording and coding responses can be reduced, and potential inter-rater variability can be minimised.

Evidence from experimental and quasi-experimental studies on aspects of questionnaire appearance was scanty. However, a number of articles were identified that outlined a theoretical basis to aspects of design, which suggested that questionnaire appearance can influence respondents' decisions at several stages, from arousal of interest in questionnaire completion, through task evaluation, to initiation and monitoring of the process of completion. There is a need for consistency in the presentation of visual information and an understanding and application of "graphic non-verbal language" (i.e. the spatial arrangement of information and other visual phenomena such as colour and brightness).

Enhancing response rates

High survey response rates are desirable because they increase the precision of parameter estimates and reduce the risk of non-response bias.

Many factors may combine to influence the decision of a recipient of a questionnaire to respond. Potential respondents must have both the means to complete the questionnaire and the will to do so; the perceived costs of responding must not exceed the benefits.

"Saliency" – the apparent relevance, importance and interest of the survey to the respondent – is a very important influence on response rates. Fortunately, health-related surveys are likely to be perceived as salient. Perhaps surprisingly, questionnaire length appears to be less important.

The number of contacts made with sampled individuals is another powerful factor. Some researchers advocate prenotification, so that recipients are primed for the arrival of the questionnaire. Almost all experts in survey design advocate the use of reminders, a recommendation supported by evidence from primary studies.

Other factors that have been shown to influence response rates include making a self-interest/utility appeal to the respondent and the use of incentives (particularly enclosed monetary incentives). Perhaps surprisingly, anonymity has not been demonstrated to have any consistent effects on the rate or quality of response.

Conclusions

Recommendations for practice

The heterogeneity of findings indicates that there can be no universal recommendations on best practice in respect of questionnaire design and survey conduct. Rather, individual survey researchers need to take into account the aims of the particular study, the population under investigation and the resources available; trade-offs between the ideal and the possible are likely to be needed. However, some general principles can be offered.

The principal objective should always be to collect reliable, valid and unbiased data from a representative sample, in a timely manner and within given resource constraints.

In choosing a mode of questionnaire administration, consideration needs to be given to the availability of an appropriate sampling frame, anticipated response

rates, the potential for bias from sources other than non-response, acceptability to the target population, the time available, the financial budget, and the availability of other resources (e.g. skills or equipment).

In formulating questions and response categories, and in determining question order, researchers should bear in mind that survey respondents employ a wide range of cognitive processes in formulating their responses. To minimise bias and to reduce spurious inter-respondent variation, careful attention must be given to these issues.

The "task analysis" model, the theory of social exchange and theories of perception and cognition should inform decisions regarding the physical design of questionnaires, as well as strategies for delivering and returning them. The aim should be to enhance the perceived and actual benefits of responding and to minimise the perceived and real costs. The effort required to interpret questions and provide responses should be made as easy as possible. Strategies for reducing the monetary cost to respondents include the use of prepaid return envelopes and the provision of financial incentives (unless ethical imperatives preclude the latter).

Recommendations for research

Both quantitative research (in the form of experimental manipulations of various aspects of questionnaire design and administration) and qualitative research (in the form of cognitive interviews addressing the processes by which respondents react to questionnaire stimuli) are required.

Assessing the reproducibility of previous findings should be afforded higher priority than embarking on totally new lines of enquiry. In particular, it will be important to investigate whether findings from social, educational or market research also apply to health-related surveys. It will also be important to test whether observed effects of manipulating different aspects of questionnaire design are equally applicable to interviewer-administered and self-completed questionnaires.

Multiple measures of outcome or "success" should be examined, including those of quantity (e.g. questionnaire and item response rates) and quality (e.g. non-response bias; and validity, reliability and distribution of responses), as well as resource implications.

Chapter 1

Background and introduction

What is quantitative survey research?

The quantitative survey method may be defined as "a set of scientific procedures for collecting information and making quantitative inferences about populations".

Within health technology assessment, health needs assessment, epidemiological research, audit and other quality initiatives, questionnaire surveys are frequently the method of choice for gathering primary quantitative data from patients and healthcare professionals. Questionnaires used in these contexts need to provide valid, reliable and unbiased data from a representative sample of respondents. Depending on the aim of the data collection exercise and on the study design, the data must be capable of one or more of the following:

- discriminating between groups and individuals at a given point in time
- detecting change over time
- predicting future behaviour or needs.

Distinguishing features of the survey method

A number of features make the survey method different from other methods of research.

- Surveys involve collecting new data, rather than being purely theoretical or based on information already available.
- Surveys involve collecting data about a population of units of some defined type, but they are usually based on samples, rather than complete censuses, of these units.
- The method for selecting the sample is fixed and objective; ideally, it should be based on statistical probability theory.
- The sample of units for which data are collected is often large (hundreds or even thousands).
- The procedures for sampling these units and for collecting information are explicit, systematic and standardised.
- There is careful prior definition of the information to be collected from each sampled unit.
- The data collected and the results are quantitative (counts, rates, etc.); measurements are applied in a standardised way to each sample unit in order to achieve objective measurement of concepts, attributes and so on, and thereby yield comparable results across the entire sample.
- The data collected are such that they can be handled purely arithmetically (quantified); qualities or attributes (e.g. gender, strength of opinion) are assigned a numerical code so that they can be more readily manipulated in statistical analyses.
- The results from measurements made on the sample are summarised statistically.
- Conclusions about the population are drawn within confidence limits defined by using sampling theory (i.e. inferences are made from the sample to the underlying population).
- Conclusions may be descriptive of the population or based on the testing of specified hypotheses.
- Surveys are conducted in the "real world", under circumstances that cannot be fully controlled, rather than in the more rarefied laboratory setting of biomedical research.
- Surveys use a wide range of human skills and other resources and require much planning and teamwork.

Quantification in surveys

Quantitative surveys aim to convert the information they collect to meaningful numbers (e.g. counts, averages, rates) relating to the population of interest. It is not good enough simply to produce numbers if their meaning is not clear or to elicit information that does not yield useful numerical estimates. Because most surveys are sample surveys (rather than population censuses), some random sampling variability in the results obtained is inevitable (i.e. if the same questionnaire were administered to a different sample drawn from the same population, the findings from the two samples would not be identical). For this reason alone, the results from sample surveys will always be estimates, surrounded by margins of error (in statistical parlance, parameter estimates with associated confidence limits). One of the main reasons for favouring

probability sampling methods is that they make it possible to estimate the margins of random sampling error. The key aims of optimised probability sampling are, therefore: to minimise random sampling variability (variance); and to avoid or minimise systematic sampling bias.

Meaningful numbers are produced by classifying, counting and scoring respondents' answers to survey questions. These procedures are followed at the analysis stage by summarising and estimating, whereby inferences are made from the sample on which the data were collected to the underlying population.

Types of information that may be gathered in a survey

Depending on the objectives of the data collection exercise, Dillman has suggested that questionnaire surveys may be used to gather one or more of the following types of information (p. 80).[1]

- Attributes – according to Dillman,[1] "what people are" (e.g. personal characteristics, such as age, gender, marital status, personal and familial medical history, educational achievements, occupational status): Attributes may be seen as something that are an intrinsic and relatively stable part of the respondent (at least at a given point in time), as opposed to something he or she does. Attributes are generally regarded as facts, but of course the reporting of facts in surveys may be subject to distortions.
- Behaviour and events – "what people do" (e.g. frequency of engaging in potentially risky behaviour such as smoking or alcohol consumption) or "what has happened in people's lives" (e.g. having a particular acute disease, suffering a bereavement): Questions about behaviour may refer to past, current or (intended) future behaviour; questions about events usually refer to the past.
- Beliefs/knowledge – "what people think is true" (e.g. beliefs and understanding regarding the causes of illness): Assessing beliefs means assessing what respondents believe to be true or false, correct or incorrect. There is no implied value judgement about what is "good" or "bad". Belief questions include those that test a respondent's knowledge of facts, as well as those that tap into issues for which there is no agreed "correct" answer.
- Attitudes/opinions/reasons etc. – "what people say they want"; "how people feel about something" (e.g. satisfaction with healthcare services): Attitudes and opinions are essentially evaluative, reflecting respondents' value judgements about what is good or bad, effective or ineffective, desirable or undesirable. Measuring attitudes involves making assumptions about how people structure their perceptual world. For example, it is no use asking questions to elicit people's attitudes to "services provided by X healthcare trust" if they have no concept of what the healthcare trust is or does. Sometimes, questions may be posed in such a way that they tap a mixture of attitudes and beliefs, but the distinction is still usefully made.[1,2]

It is important to bear these distinctions in mind when designing questionnaires and determining question wording. Unless the survey researcher is clear about what is to be measured, the information yielded by the question may not be what is actually required.

Quality aims in survey research

The key aims in quantitative social surveys are to collect information that is:

- valid: measures the quantity or concept that is supposed to be measured[3]
- reliable: measures the quantity or concept in a consistent or reproducible manner[3]
- unbiased: measures the quantity or concept in a way that does not systematically under- or overestimate the true value[4]
- discriminating: can distinguish adequately between respondents for whom the underlying level of the quantity or concept is different.

The risk of collecting information that fails one of these tests is ever present.

Why this review was needed

Close attention to issues of questionnaire design and survey administration can reduce these errors and bias. However, few healthcare professionals have any formal training in data collection or questionnaire development. Even among the academic and research communities there is considerable reliance on tradition and conventional wisdom, rather than on sound theories of respondent behaviour and evidence from empirical studies. A number of the classic texts[1,5–7] on questionnaire development, on which many researchers and health surveyors rely, are now quite dated. Moreover, some of these texts

draw mainly on the accumulated experience and opinions of the expert authors, rather than on evidence from experimental studies. Although Bradburn and Sudman[8] found that "for the most part we were gratified to find that the data [from experimental manipulation of questionnaire wording and administration methods] confirmed our intuitions", there were "just enough surprises to warn us that we should not rely entirely on our own experience but should check things out empirically whenever possible". Furthermore, texts aimed at a particular discipline[6,9] may fail to reflect recent developments and current best practice in research methods by relying solely on findings from health services and epidemiological research, rather than learning from developments in psychological, social and market research.

Target audience for the review

Although the research brief for this review was the use of patient and staff (i.e. health professional) questionnaires in the context of health technology assessment, the authors recognise that questionnaires are widely used in other fields of health research, such as epidemiology, and, of course, the use of questionnaires and survey methods in general is not confined to the health sector. Indeed, the literature review draws widely on articles from the fields of social, educational and market research. However, in synthesising the findings, distinction has been made between evidence from studies on health-related topics and those on other subjects and of other populations. Similarly, in interpreting the findings and making recommendations for practice, the authors have tried to take account of the particular features of health-related research (e.g. ethical concerns) and to indicate the extent to which findings from other sectors are likely to be generalisable to this particular sector.

The immediate target audience, therefore, is researchers and practitioners using questionnaires in assessing "health technologies", defined as:[10]

> "all the methods used by health professionals to promote health, to prevent and treat disease, and to improve rehabilitation and long-term care. These methods include 'hardware' such as syringes, medicines and high technology diagnostic imaging equipment; 'software' such as health education, diagnostic and therapeutic policies, as well as the skills and time of people working in the health services."

However, the requirements of validity, reliability and feasibility in data collection apply equally to other users of questionnaires in the health sector, such as in health needs assessment and other epidemiological research, and in audit and other quality assurance initiatives. The authors therefore believe that such users will also find much of relevance and interest to them in this review.

Aim and objectives of the review

The principal aim of this review was to address the evidence gap identified above. The key objectives were:

- to identify established and innovative approaches to questionnaire design and administration and thereby to identify current "best practice" with respect to the design and conduct of questionnaire surveys (by "best practice" is meant practice informed by sound theory and empirical evidence from well-designed studies, as well as the accumulated experience and opinions of expert practitioners in survey research)
- to identify, analyse and synthesise evidence for ways in which the quality of survey data (particularly validity, reliability and lack of bias) may be improved by attention to aspects of questionnaire design and survey conduct
- to identify practical issues surrounding questionnaire design and survey conduct, particularly with respect to resource implications
- to evaluate the extent to which approaches from other disciplines are likely to be transferable to a health-specific context and in particular to health technology assessment
- to identify gaps in current knowledge and hence to recommend topics for further research into issues of questionnaire design and survey conduct.

Scope of the review

An operational definition of "questionnaire"

The term "questionnaire" has been used to describe a variety of data collection instruments. For example, Franklin and Osborne[11] defined a questionnaire as "an instrument consisting of a series of questions and/or attitude–opinion statements designed to elicit responses which can be converted into measures of the variable under investigation" (cited by Nay-Brock[12]).

For the purposes of this review, the term is defined to mean "structured schedules used to elicit predominantly quantitative information, by means

of direct questions, from informants, either by self-completion or via interview".

Topic guides for use in semistructured or unstructured qualitative interviews were excluded from the definition, as were schedules or proformas for the primary collection of observational or clinical data, or for the abstraction of secondary data from sources such as medical records.

The inclusion of schedules for use in structured interviews was in recognition of the fact that interviewer administration (whether face-to-face or by telephone) may be the most appropriate method of data collection in certain circumstances (see chapter 3). However, the authors recognise that self-completion questionnaires are likely to be the method of choice in many health surveys, a choice often dictated by resource constraints. In presenting the evidence from primary studies, the mode of survey administration has been highlighted. In interpreting and synthesising the findings, the extent to which results of interview surveys can be extrapolated to self-completion questionnaires and vice versa was considered.

Included aspects of questionnaire design and survey conduct

Most basic texts on survey methods are in general agreement regarding the steps involved in carrying out a survey. *Box 1* summarises the steps as set out by Oppenheim.[13] The shaded areas in this box define the scope of the body of this review; the starred topics are dealt with briefly in appendix 1. In drawing the boundaries to the scope of the review, the focus was on those aspects of the survey process that are most amenable to experimental manipulation and generalisation.

BOX 1 Steps in a survey (after Oppenheim[13])

- Define the aims of the study
- Review the current state of knowledge on the topic
- Conceptualise the study
- Determine an appropriate study design (e.g. experimental vs. observational, prospective vs. retrospective) and assess feasibility within resource constraints
- Decide upon hypotheses to be investigated, determine and operationalise data requirements
- Choose the most appropriate method of data collection (e.g. self-completed questionnaires vs. interviews)
- Design or adapt data collection instruments
- Conduct pilot work and refine methods and instruments*
- Design and select sample*
- Conduct data collection (often termed "field work")
- Process data
- Analyse data
- Report findings

The principal foci of the review were therefore:

- choice of mode of survey administration (face-to-face and telephone interviews; postal and "captive audience" self-completion questionnaires)
- methods of recording responses (traditional "pencil-and-paper" techniques; computer-assisted techniques)
- issues of question wording, choice of response formats, and question sequencing
- issues of questionnaire formatting and other aspects of layout
- survey administration, with special emphasis on postal surveys, in particular to enhance response rates and reduce threats of bias.

Appendix 1 provides some brief guidance on aspects of the survey process that are not included in the review proper; this guidance is based primarily on the collective experience and practice of the authors of this report, although references to some key texts and articles are also provided. For further guidance on the steps beyond the scope of this review, the reader is referred to the many comprehensive texts available.[1,3,5,9,13–21] Readers who are interested in developing measurement scales should refer to Streiner and Norman.[22]

Framework for presentation and appraisal of evidence

As noted above, the primary objective in survey research is the collection of valid and reliable data. At each stage in the survey process, threats to validity arise and there is the potential for bias to be introduced. Bias has been defined by Sackett,[4] following Murphy,[23] as "any process at any stage of inference which tends to produce results or conclusions that differ systematically from the truth".

These threats to data quality, and the methods by which they can be minimised, are highlighted in the chapters that follow. However, researchers and health surveyors generally operate within limitations of time, money and other resources, and also within ethical boundaries. Often, a trade-off is required between what is optimal in terms of data quality,[24,25] and what is practicable in the face of such constraints. In the presentation of opinions, findings and recommendations, issues of timeliness, cost and other resource implications are also highlighted.

Structure of this report

In chapter 2, the methods used in this review are presented, with critical consideration of the limitations imposed by those methods, together with the implications of the approach for users of this review. In chapter 3, modes of survey administration are discussed. Issues of question construction and sequencing are considered in chapter 4. Chapter 5 is concerned with aspects of questionnaire appearance and layout. In chapter 6 the importance of high response rates is considered and how these can be encouraged is discussed, while giving due concern to other aspects of data quality. Chapters 3–6 follow a common structure: first is an exposition of "expert opinion" as encapsulated in textbooks on survey methods; where appropriate, relevant theoretical perspectives (e.g. theories of respondent behaviour) are also presented; next, the evidence from primary research studies is reviewed; this review is followed by a summary of conclusions from the available evidence; finally, each chapter concludes with a series of recommendations for "best practice" and for future research. For convenience, the conclusions, recommendations for practice, and recommendations for future research are all drawn together in chapters 7, 8 and 9 respectively. Finally, in chapter 10, the trajectory of the knowledge base on the topic of survey design and administration is briefly considered.

In appendix 1, non-evidence-based guidance for key aspects of the survey process that have been excluded from the review proper are presented. Appendices 2–4 contain the documents used in conducting the review: a list of topics for data abstraction; the data abstraction form; and the data abstraction manual. Appendix 5 provides a list of articles that meet the inclusion criteria and pass the methodological screen, but were inadvertently omitted from the review proper.

Chapter 2

Methods of the review

In this chapter, we set out the methods used in this selective, narrative review of the literature on questionnaire design and survey administration. We consider the limitations of our chosen methods, with particular emphasis on threats to validity and generalisability arising from the constraints we imposed upon ourselves and had imposed on us. We conclude with the implications of our approach and of the limitations for users of this review.

Defining the scope of the review

Rationale for a selective, narrative review

Writing in 1974, Sudman and Bradburn[26] cited 935 references on the topic of error in surveys. In a bibliography of market research compiled in 1986, Dickinson[27] cited 454 studies of mail survey responses. An ongoing systematic review of methods to influence response rates to postal questionnaires has identified 282 published randomised controlled trials (Edwards P, Institute of Child Health, University College, London: personal communication, 2001). From the outset, we recognised that, because of resource constraints, a comprehensive review of all the published literature (even of all the randomised trials) on the subject of survey methodology would not be feasible. Indeed, such an exhaustive review would not even be desirable because literature published in the 1950s, for example, is unlikely to be of great relevance today. This recognition was echoed by a number of researchers undertaking "systematic reviews" of methodologies for health technology assessment.[28,29]

We also recognised from the outset that the highly rigorous approach adopted by the Cochrane Collaboration[30] in their systematic reviews of healthcare interventions was not practicable or appropriate to this review. Cochrane Collaboration reviews require that the review protocol (search terms; inclusion, exclusion and quality criteria; outcome measures) must be defined *a priori* and should not be revised once the review in under way. Reviews on methodological topics, however, frequently require an "iterative" approach because the topics are wide-ranging and often have ill-defined boundaries.[29] This was indeed our experience; the sheer volume of literature available, the breadth of the topic and the potentially fuzzy boundaries forced us to refine and redefine our scope as the review got under way.

Nonetheless, in identifying the imperative to be selective about what literature we included, we also recognised the need to be explicit about the limits we set, the methods we used and the potential limitations imposed by our choices, and to impose a consistent structure on how identified studies were appraised, synthesised and reported. By indicating which databases were searched and which search terms were used, and by making explicit our inclusion and exclusion criteria and the means by which findings from identified papers were analysed and synthesised, we believe that other researchers "using the same standards and methods, should detect the same literature and come to the same factual conclusions" (p. 250).[28] In this respect, the "processes by which the literature were obtained and synthesised [were] at once methodical and explicit".[29]

We acknowledge that our decisions and actions regarding the scope and coverage of this review mean that it must be regarded as a selective, narrative review rather than a systematic review. However, we believe that our explicitness regarding inclusions and omissions, and the likely implications of our decisions and actions, will allow readers to judge whether the findings and recommendations reported here need to be supplemented by a search for further confirming or refuting evidence.

Defining the boundaries

In setting the boundaries for our literature search, we were guided primarily by the aims and objectives set out in chapter 1. We also took into account the subject matter of the other reviews being undertaken under the NHS HTA Methodology initiative and sought to avoid significant overlap with the work of other methodology reviewers.

Our key question in determining whether an identified study fell within the scope of this review was: Does this work (appear to) help to identify best practice with respect to survey design and administration? If the answer to this question was

"yes", a hard copy of the article was sought, read and, if appropriate, abstracted and synthesised. To aid this decision further, we developed a set of explicit inclusion and exclusion criteria (set out below) and a topic list (appendix 2) to which reviewers referred in scanning abstracts and articles.

The inclusion criteria and topic list were derived from our stated objectives (chapter 1); we went through an iterative process of refining these criteria during the early stages of literature searching to reflect issues raised by reviewers, particularly dilemmas in determining whether studies identified by our search strategies should be included or not. In defining many of our exclusion criteria, we set out explicit justification for doing so.

Final inclusion criteria:

- comparison of questionnaires/interviews versus diary methods, unless concerned solely with the validity of the data yielded by the competing approaches
- studies of particular approaches to structured interviewing (e.g. computer-assisted interviewing), if these approaches were being used in the context of research as opposed to clinical practice
- articles reporting the context (e.g. patient groups, subject matter) in which interactive computerised techniques have been employed, since this is a novel topic
- articles reporting the context in which video or audio questionnaires have been employed, again because of the novelty of the topic
- studies of particular types of question (e.g. situational response, circular questions), particularly if these were novel
- studies of question ordering (e.g. relative position of general and specific items)
- comparison of short-form and long-form measures or questionnaires, but only if the focus of the article was the effect on response rates or other key issues identified in the topic list, not if the focus was solely on the reliability/validity of the information yielded
- aspects of questionnaire formatting, including size, colour and print styles
- techniques for enhancing response rates, especially in postal and other self-completion surveys.

Exclusion criteria:

- Studies of qualitative approaches to data collection (e.g. unstructured interviews, focus group discussions): We had defined questionnaires to mean "structured schedules used to elicit predominantly quantitative information, by means of direct questions, from informants, either by self-completion or via interview" (chapter 1). Moreover, qualitative approaches formed the focus of another methodological review funded by the HTA Programme.[31]
- Comparisons of structured (questionnaire) approaches with unstructured (qualitative) approaches: Again, this exclusion was based on our decision to focus on the collection of predominantly quantitative information.
- Studies addressing the development or refinement of scales to assess attitudes or opinions, or to measure health status or quality of life: Although standardised health status/quality of life measures are often included in questionnaires administered to patients, we argued that this subject was already fairly well covered in the literature, for example by Streiner and Norman.[22]
- Studies (whether comparative or not) assessing the practical or psychometric properties of specific health status/quality of life measures: Again, we considered that these issues were already fairly well covered in the literature.[32–35] Moreover, another methodological review funded by the HTA Programme was addressing certain aspects of this topic.[36]
- Articles reporting techniques for establishing reliability, validity or responsiveness to change: This was never intended to be a review of psychometric principles and methods.
- Studies focusing on use of interviews/questionnaires for clinical purposes (e.g. screening/health history taking/counselling); likewise papers reporting on methods for training clinical interviewers or improving their performance: We added this criterion when we found that the MeSH term "interviews" in the MEDLINE database and the thesaurus term "interviews" in PsycLIT were not specific to interviews used in the context of survey data collection.
- Job, selection and media interviews: Again, this criterion was added because of the lack of specificity of the search term "interviews".
- Studies of the use of questionnaires specifically in the context of Delphi surveys or other consensus methods: This criterion was based on the fact that consensus methods formed the focus of another methodological review funded by the HTA Programme.[37]
- Articles that simply reported on the use of questionnaires/interviews without any comparison or evaluation of the effectiveness of the approach, unless they were concerned with novel techniques

such as interactive computer-assisted questionnaires or video questionnaires (see above): Setting this limit was essential in keeping the number of identified publications to a manageable number. Furthermore, we did not believe that such general publications would offer any useful insights into best practice in questionnaire design and administration.
- Articles reporting solely on the context in which telephone interviews have been employed: Telephone interviews are not a novel topic.
- Articles solely reporting on the context in which non-interactive computerised techniques have been employed: This too is no longer a novel topic.
- Comparisons of interview/questionnaire approaches with objective assessments (e.g. clinical examination, record-based approaches): Although a comparison of subjective self-reports with objective (in theory, at least) sources is frequently a means of validating the former, such comparisons are generally topic- and population-specific and therefore are not readily generalisable.
- Studies of the applicability of the questionnaire approach to particular topics (e.g. assessment of diet or smoking behaviour): Although we recognised that this is a very important issue for survey researchers, defining limits to the potential topics would have been almost impossible. (This is a good example of the "ill-defined boundaries" referred to by Edwards and colleagues.[29]) We recommend, therefore, that researchers who are contemplating the use of a questionnaire survey to gather information on a particular topic should first conduct their own literature review on the likely "yield" (in terms of both quantity and quality of data).
- Studies of the applicability of the questionnaire approach, or the appropriateness of different survey methods (e.g. face-to-face interviews versus self-completion questionnaires) for particular respondent groups (e.g. elderly, homeless or mentally ill people): Our reasons here were the same as those leading to the exclusion of the applicability of the questionnaire approach to particular topics. Once again, we recommend that researchers should carry out their own literature review on the appropriateness of surveys in general, or on particular modes of survey data collection, for their population of interest.
- Studies on the impact of interviewer characteristics (e.g. age, gender, ethnicity) on response rates and response quality: We initially intended to include this issue but early in the review process it became clear that such effects are population- and topic-specific. We highlight this fact in chapter 3 and once again recommend that researchers should check out the relevant literature pertaining to their target population and topic of interest.
- Comparison of proxies versus self-report: We initially intended to include comparative studies of proxy- and self-report where the focus of the article was the effect on response rates or other key issues identified in the topic list, but not those with a focus solely on the reliability/validity of the information yielded. However, early in the review process it became clear that such effects are population- and topic-specific and we therefore dropped this topic. We recommend, however, that researchers who are contemplating the use of proxy informants should review the relevant literature for their own population and topic.
- Articles reporting on the impact of questionnaires on respondents (e.g. whether they generate anxiety): Such impacts are likely to be context- or topic-specific and general messages on "best practice" are not likely to be available.
- Comparison of short-form versus long-form measures, if the focus of the article was solely on the psychometric properties (validity and reliability) of the two versions.
- *Post hoc* comparisons of respondents and non-respondents to specific surveys: Although there may be some general trends with respect to who does and who does not respond, this information is not likely to inform best practice. However, outputs from the review (see in particular chapters 3 and 6) mention the likelihood of non-response bias and remind readers to anticipate it and test for it.
- Studies of research ethics in general terms: Ethical issues are likely to be study specific and should therefore be reviewed on a study-by-study basis. Moreover, general ethical issues formed the subject matter of another methodological review funded by the HTA Programme.[38]
- Sampling methods: Although selecting the sample is a key part of any survey, we considered that issues to do with sampling – choice of sampling frames, sampling techniques (e.g. probability versus non-probability; random versus systematic) and calculation of sample size – were all likely to be study specific; we therefore thought that it would be difficult to recommend universal "best practice". However, in appendix 1, we highlight some key issues to consider in selecting a sample.
- Selection and application of previously developed questions and scales: It would be almost impossible to define boundaries to the topics that could conceivably be covered in

surveys of healthcare consumers, patients or healthcare professionals (e.g. they may well include aspects of educational experience, job satisfaction etc., as well as health-related issues). However, in appendix 1, we provide some pointers to major sources of existing questions and scales.
- Piloting of questions and questionnaires: Again, this is a major topic in its own right. However, we recognise the importance of adequate piloting of questionnaires and survey protocols, and highlight this at appropriate points in the review (particularly in chapter 4); we return briefly to this issue in appendix 1.
- Interviewer training, also as a major topic in its own right: In chapters 3 and 6 we refer to the importance of adequate training, but we do not offer detailed advice on its provision. In appendix 1 we provide some references to further reading on this topic.
- Use of census data in coding occupations: This is a highly specialised topic and unlikely to be of common concern to the target audience.
- Data checking and cleaning techniques: This is a major topic in its own right and is related to issues of data analysis.

Topics excluded by default

There were a few topics that we did not deliberately or explicitly exclude from our review, but upon which no relevant comparative publications were identified by our literature search. Resource constraints precluded us from a more intensive search for literature on these topics:

- individual tailoring of questionnaires to respondents
- effects of revealing researchers' bias
- techniques for handling sensitive topics (e.g. random response).

Language and location

Resource constraints led to the decision to confine our review to articles published in English. However, we did not confine it to studies carried out in the UK; indeed, no geographical constraints were imposed. We recognise that cultural differences between, for example, North America and the UK may limit the generalisability of some findings (e.g. postal systems and rates differ between the USA and the UK). However, much innovative work in survey design and administration has been carried out outwith the UK; we believed that it was very important to include such work. Nonetheless, in presenting conclusions and recommendations, we advise caution in overgeneralising from one culture to another.

Period of study

Resource constraints also dictated that we impose a time cut-off. A number of the classic texts on questionnaire design and survey methodology date from the 1970s.[1,5] We also perceived that that period also marked a significant growth of interest in the use of patient and staff questionnaires for outcome assessment. We therefore decided to take 1975 as the starting point for our review and to seek evidence from studies published in that year or later (with an upper cut-off of 1996). Only if there was insufficient evidence from 1975 or later, would evidence from earlier studies be sought; however, in pursuing secondary references (see below) in respect of theoretical aspects of survey research, we deemed it appropriate to access seminal work from earlier years.

Disciplines

A key decision was to include evidence from fields other than health. Resources for our review were limited and, by confining our search to evidence from health-related questionnaires and surveys, we could have extended the review to further databases and could perhaps have carried out hand-searches of key journals. However, we recognised that survey research is not confined to health researchers. Psychologists and social and market researchers have led many of the developments in this field. Theories of respondent behaviour derive mainly from psychologists. Methodological messages are, in the main, unlikely to be discipline specific. We therefore considered that it was appropriate to seek evidence from studies published in journals primarily devoted to social and market research and to psychology (e.g. *Public Opinion Quarterly, Journal of Applied Psychology, Journal of Marketing, Journal of Marketing Research*) as well as from those published in the medical and health-related literature. However, we recognised that it would be important to highlight the sector and topic of each survey reported, in order to allow judgements to be made about generalisability. In drawing conclusions and making recommendations for practice, we note the need for caution in generalising from non-health-related work to health surveys.

Types of study/levels of evidence

Although the randomised controlled trial is generally considered to be the "gold standard" method for evaluation of interventions,[39] we recognised that, in the absence of well-designed trials, useful information may be gleaned from less rigorous study designs. Moreover, we knew that useful knowledge about which techniques work well is often garnered through long years of experience and that an understanding of theories

of respondent behaviour provides a useful basis for recommendations regarding the conduct of surveys. As Edwards and colleagues state, in contrast to reviews of clinical interventions, reviews on methodological topics do not have a "gold standard" against which different methods may be judged; such reviews (this one included) are therefore "likely to rely on extracting argument (perhaps as well as data) from the literature" (p. 255).[29]

We therefore believed that it would be inappropriate to confine our review solely to evidence from experimental studies and decided to include:

- articles reporting on experimental designs (in particular, randomised controlled trials) in which two or more approaches to questionnaire design and/or administration were manipulated and compared
- articles reporting on other comparative studies (usually using quasi-experimental designs, such as non-random concurrent controls or historical controls) in which two or more approaches to questionnaire design and/or administration were compared
- articles reporting on empirical studies, in which the advantages and disadvantages of a single approach to questionnaire design and/or administration were reported (primarily where such reports related to a novel technique such as audio questionnaires)
- articles propounding theories of respondent behaviour
- previous review articles (as a source of useful references, as well as a summary of previously gathered evidence).

Nonetheless, as described in greater detail below, in synthesising results from empirical studies we prioritised findings from randomised controlled trials as providing the highest grade of evidence.[30]

Modifications to proposed approach

As indicated above, a number of modifications to the originally proposed approach were made during the execution of the review. The inclusion and exclusion criteria set out above were not defined once and for all at the beginning of the study. Instead, they were developed and refined in an iterative process as the study progressed, in the light of our experience. For example, our initial scan of the literature indicated that the issue of whether a questionnaire is a valid and appropriate means of data collection, or whether an objective assessment is required, is context- and topic-specific. As indicated above, we therefore decided to exclude studies on the applicability of the questionnaire approach to particular topics (e.g. assessment of diet or smoking behaviour). Similarly, we agreed to exclude studies of the applicability of survey methods in collecting data from particular subject groups (e.g. elderly, homeless or mentally ill people). Nevertheless, we recognised that these issues need to be considered at the beginning of any data collection exercise. We therefore advise researchers to carry out a population- and topic-specific review of the appropriateness or otherwise of a subjective questionnaire approach, to inform their choice of method of data collection.

We had originally proposed to include issues of data preparation and validation in this review. However, on more detailed consideration, we considered that these issues were likely to be highly context specific and also specific to the software packages used for data entry, validation and analysis.

As noted above, we did not identify any published studies on some of the areas defined by our inclusion criteria. For example, our initial search did not identify any publications on the effects of revealing researchers' biases to respondents. Within our resource constraints we were unable to search further for relevant material.

Search strategy

Proposed approach

Key texts

Most researchers do not approach survey data collection *ab initio*, but rather draw on "conventional wisdom", often encapsulated in existing textbooks or in what they have been taught at undergraduate or postgraduate level. We thought, therefore, that it would be appropriate to preface our review of empirical studies with a summary of current "expert opinion" and then to seek confirming or refuting evidence from primary research. Sources of expert opinion were key texts on survey methods (e.g.[1,5,13,40]). These were identified through the personal and institutional libraries of the research team.

Databases

In our search for primary research studies, previous literature reviews and theoretical articles, we initially prioritised four electronic databases for searching:

- MEDLINE
- PsycLIT
- CINAHL
- EMBASE.

MEDLINE, CINAHL and EMBASE were prioritised because of their health-related focus; PsycLIT was selected because we knew that much of the experimental work on questionnaire design and administration is published in the psychological, sociological and market research literature indexed on that database. Among these four databases we further prioritised MEDLINE and PsycLIT because we anticipated that these would provide the highest yield.

We proposed that the search would subsequently be extended if insufficient good-quality evidence was found from these initial databases. In order of priority, other databases to be considered would be: ASSIA Plus, Social Science Citation Index, BIDS-ISI, Educational Resources Information Centre, British Library Holdings, Business Periodicals Index, and Index of Theses and Dissertations.

Search terms

Initial scanning of the MEDLINE and PsycLIT databases indicated that detailed methodological keywords were inconsistently applied. Rather than searching for specific terms such as "incentives", we opted for a simple keyword search strategy (MeSH headings in MEDLINE; Thesaurus terms in PsycLIT) using the terms:

- surveys (PsycLIT only)
- questionnaires
- interviews.

(We initially considered searching for these terms in titles and abstracts as well as in MeSH headings and thesaurus terms, but found that such a broad search strategy was of extremely low specificity, yielding thousands of articles simply citing questionnaires as the mode of data collection.)

Obtaining further references

We also proposed that the reference lists of articles identified through the search strategies described above would be scanned to identify other potentially useful references (within our data span of 1975–1996) in a citation pearl-growing approach. We also proposed to adopt a citation index approach, whereby electronic citation indexes would be used to identify other publications citing key references previously identified through the search strategy described above.

Finally, we agreed that, if the yield from other strategies was low and time permitted, the Index of Theses and Dissertations (UK and international) and the British Library Holdings would be searched. Furthermore, we considered hand-searching key high-yield journals such as *Survey Methods Bulletin* and *Public Opinion Quarterly*.

Sifting and selecting references

Our proposal provided for a multistage sift.

Stage 1: identification of potentially relevant articles

The initial sift of included studies (on the basis of our explicit inclusion and exclusion criteria) was on the basis of title, abstract (where available) and subject headings (keywords). Reviewers worked either directly from the screen-based electronic record retrieved from the database or from a printout of the records retrieved by the electronic search strategy, according to personal preference. Each identified article was initially categorised as "definitely include", "possibly include" or "do not include". If the title and abstract did not provide sufficient detail to make a definitive decision, reviewers were advised to err on the side of overinclusiveness at this stage and seek a copy of the full article.

The initial stage of the sift was carried out by two reviewers per database until an inter-reviewer agreement level of 80% was attained. At that point, responsibility for the initial sift of that database was divided amongst the two reviewers allocated to that database (on a year-by-year basis). When doubts arose about whether a particular article should be included or not, the reviewer consulted at least one other member of the review team.

We sought copies of all articles identified as "definite" or "possible" inclusions, if necessary through the British Library's Inter Library Loans system. All references identified in this way were added to an electronic database using the Reference Manager software.

Stage 2: reassessment of identified articles

Ideally, in a formal literature review, each article should be assessed independently by two reviewers. However, resource constraints precluded this. The second stage of the sift was therefore carried out by five reviewers working independently, once all five had worked through a "training set" of articles to establish inter-reviewer consistency (not formally quantified).

Articles were sorted into batches of 30, based on the Reference Manager identification number, and allocated to reviewers on a batch basis. Each batch could include both articles that that reviewer had screened at the first sift and those initially screened by other reviewers. This approach was adopted to minimise the risk of selection bias in

favour of articles previously identified by a given reviewer as "relevant".

Each article was read and reassessed against the inclusion and exclusion criteria. We key-worded all those meeting the inclusion criteria. For all comparative studies (randomised controlled trials or quasi-experiments) we applied a 5-point methodological screen (see appendix 3):

1. minimum group sizes ≥50 at the time of allocation
2. participants randomly allocated to groups (studies in which the entire sample was selected randomly but allocation to groups was subsequently quasi-random or systematic were also scored "yes" on this criterion)
3. control and intervention groups comparable at baseline
4. methodological intervention clearly stated
5. methodological intervention formally evaluated.

In order to be included in the quantitative review, an answer of "yes" to the first, fourth and fifth criteria, and to at least one of the second and third criteria, was required. Full details of studies meeting these quality criteria were abstracted.

Modifications to proposed approach

Resource constraints and limitations imposed by the University of Newcastle library and the British Library caused us to modify our proposed approach. The number of articles identified led the review team to exceed their normal allocation of interlibrary loans. Although we were able to negotiate an increase in this allocation, each additional article requested through this system was charged at £1.50 rather than the original price of 30 pence. A limit of six loans per day was also imposed. Reading an article and assessing it against inclusion and exclusion criteria took a minimum of 5 minutes; key-wording and data abstraction took anything from 5 minutes to 1 hour, depending on the length and complexity of the article.

The very high yield (almost 700 articles) from the searches of MEDLINE and PsycLIT, coupled with the limited resources available to us, led us to curtail our searching of electronic databases at that point; we did not proceed to search CINAHL, EMBASE or any of the other electronic databases. The resource constraints described also precluded our following up of all secondary references. Nor did we have the resources to carry out the proposed citation index searches or to handsearch journals proactively. However, we did identify a number of references through routine journal reading.

Finally, owing to human error, identified references from MEDLINE for 1987–1992 and from PsycLIT for 1979, 1991 and 1993–1996 were inadvertently omitted from the Reference Manager database. We therefore failed to obtain hard copies of these references; as a result, they were not included in the second sift and were not key-worded or abstracted. This error was not identified in sufficient time to be able to rectify it. A list of articles meeting our inclusion criteria but missed in the original search (i.e. those identified through MEDLINE and PsycLIT for the years noted above) is given in appendix 5.

Data abstraction

Proposed approach

We developed a structured data abstraction form (appendix 3) and an accompanying manual (appendix 4) for abstracting information from articles meeting our inclusion criteria. The form was divided into a number of sections. Information from this form was input into an Access database to facilitate subsequent data retrieval.

Report identification and key characteristics

In this section we recorded the identification number of the article (taken from the Reference Manager database) and the title, to facilitate subsequent retrieval. The identity of the reviewer was also noted. We also recorded whether the focus of the study was health or non-health related. Finally, we documented what type of study was described in the article (further guidance on how study designs should be categorised was appended to the data abstraction manual):

- randomised controlled trial (1)
- non-random concurrent controlled study (2a)
- self-controlled study (2b)
- historically controlled study (2c)
- cross-sectional study (3a)
- cohort study (3b)
- case-control study (3c)
- meta-analysis with systematic review (4a)
- systematic review without meta-analysis (4b)
- meta-analysis with non-systematic review (4c)
- non-systematic review without meta-analysis (4d)
- theoretical article (5a)
- position paper (5b).

Keywords

On the second section of the form, reviewers indicated which topics in our topic list (appendix 2) were discussed in the article, by circling the appropriate numbers.

Recommendations for future research

If the author(s) of the article under review had made explicit suggestions for future research, we recorded these verbatim.

Methodological criteria for inclusion as evidence

Higher grade evidence[30,39] is provided by randomised controlled trials and other quasi-experimental designs (designs 1–2c in our schema), so we decided to prioritise such studies for full data abstraction. If a study design had been coded as 3a–5b in section 1, no further data abstraction was carried out.

As outlined above, in the case of randomised trials and other quasi-experimental designs (1–2c), we initially assessed the study against a set of five methodological criteria (minimum group size ≤50 at the outset; participants randomly allocated to groups; control and intervention groups comparable at baseline; methodological intervention (explicitly) stated; methodological intervention (explicitly) evaluated).

Rating of study quality

In line with recommendations for systematic reviews,[41–43] for the purposes of reporting findings we rated the quality of each included study against five explicit criteria:

1. methodological intervention explicitly stated
2. sample size based on explicit power calculation
3. inclusion criteria for participants explicitly stated
4. factors other than those experimentally manipulated held constant (or, in the case of multifactorial experiments, balanced)
5. data presented in sufficient detail to allow the calculation of relative risks (RRs) (or, in the case of continuous variables, mean differences) and associated 95% confidence intervals (CIs).

A score of 1 was allocated for each criterion met. In the case of the fifth criterion, a score of 0.5 was allocated if it was possible to make the calculations for some but not all of the reported results. Thus the quality score could range (in theory) between 0 and 5, with a higher score denoting a better-quality study.

Summary of study

This was a key section of the data abstraction form. We first recorded details of the study design under the following headings:

- methodological interventions (i.e. which aspects of questionnaire design or administration were examined)
- setting (e.g. hospital, community etc.)
- country
- study population: group sizes, inclusion criteria, exclusion criteria
- primary outcome measures: instrument (i.e. questionnaire) response rates; instrument completion rates; item response rates; response bias; scores on specified scales; financial costs; others.

Next, the reviewer summarised the key findings of the article, under structured headings defined by the primary outcome measures described above. This summary was followed by conclusions and recommendations for practice, as stated by the authors of the article.

Administrative details

Finally, there was provision to record whether secondary references (i.e. references to apparently relevant articles published within our date span of 1975–1996) had been highlighted, if these references were already on our Reference Manager database or needed to be obtained, and whether the information from the data abstraction form had been added to the Access database.

Modifications to the proposed approach

We made few modifications to this aspect of the study. However, to avoid transcription errors in those cases where very detailed results were presented in an article, we simply recorded a reference to the page(s) and/or table(s) in which those results were presented. In addition, to avoid transcription errors, we did not add the summaries of results for any articles to the Access database.

Data synthesis

Proposed approach

In compiling the findings from our review, we identified four broad topics (corresponding basically to the foci identified in chapter 1) for which we wished to synthesise expert opinion and evidence and to make recommendations for "best practice":

- mode of survey administration (including methods of data capture and recording)
- question and response category wording and sequencing
- questionnaire appearance
- methods of enhancing response rates, with particular emphasis on postal surveys.

Each topic (and therefore chapter of the report) was allocated to a member of the review team

(AJ, EMcC, JS, LT: two reviewers worked together on mode of administration).

The synthesis began with a summary of "expert opinion", as described in key texts on survey methods.[1,5,7,17] Where relevant, theoretical perspectives (e.g. theories of respondent behaviour) had been identified in the course of our review. We summarised these theories and, for those aspects of questionnaire design and administration on which earlier literature reviews had been carried out, we summarised the key findings from previous reviews.

Having encapsulated "expert opinion" in this way, we then turned to the evidence from primary studies. Key details were summarised in tabular form, under the headings:

- author(s) and date
- study design (e.g. randomised controlled trial)
- focus of survey (i.e. health or non-health related)
- topic of survey
- country
- respondents (e.g. general population, patients etc.)
- mode of administration (i.e. interview or postal questionnaire)
- factors manipulated (i.e. which aspects of questionnaire design and/or administration were studied)
- sample sizes (for intervention and control groups)
- criteria for comparison (i.e. on what basis were the different "treatments" compared; e.g. response rates; percentage of respondents answering in a particular way; item non-response rates)
- main findings (summary of findings, quantified and, if possible, with RRs and associated 95% CIs attached).

(In chapter 6, the presentation is varied slightly. The emphasis of the entire chapter is on the enhancement of response rates, so these primary outcomes are separated from any secondary outcomes considered (e.g. non-response bias; item non-response rates; response bias; speed of response; cost). In the tabular presentation we concentrate on response rates and associated RRs; findings in respect of secondary outcomes are confined to the accompanying text.)

In the text we describe the findings in some detail and summarised the balance of evidence (e.g. total number of studies examining the effect of incentives on response rates and number showing a positive effect). We highlighted whether the evidence supported or refuted "expert opinion".

At the end of each chapter we drew together a summary of the evidence from primary studies. We then drew out recommendations for "best practice". In making these, we separated recommendations made on the basis of evidence from one or more high-quality primary studies from those based on expert opinion, previous literature reviews, theories of respondent behaviour, and the accumulated experience of the review team with respect to the conduct of surveys. When findings from primary research studies were negative or equivocal we indicated that our recommendations were derived from these findings, rather than being directly based upon them. Where evidence from primary studies was reinforced by expert opinion or previous literature reviews, or was underpinned by theory, we highlighted this fact.

Modifications to proposed approach

Synthesising and reporting the findings from our review was an iterative process. We refined and modified our approach on the basis of our experiences in producing early drafts of chapters, on feedback from all members of the review team on those early drafts, and on feedback from peer reviewers appointed by the HTA Programme.

In our original study protocol we had stated that we would seek evidence from cross-sectional, cohort and case-control studies if no higher-grade evidence (i.e. from randomised trials or quasi-experimental designs) was available. However, in practice we found that high-grade evidence was available for almost all of our key topics; where such evidence was lacking there was also a lack of relevant lower-grade evidence.

We originally intended simply to report findings as presented by the authors of identified articles. However, in many cases, the authors did not report actual statistical significance, but simply made such statements as "the use of incentives significantly increased response rates". Furthermore, in many studies in which multifactorial study designs had been used, authors failed to present summary statistics for each main effect. In both of these circumstances we carried out further analysis of the data presented whenever possible. In one case, this re-analysis led us to different conclusions from those of the original authors, a point highlighted at the appropriate point in the text.

Wherever possible we calculated RRs and associated 95% CIs for differences in percentages (e.g. response rates), and 95% CIs for differences in means (e.g. of response scores).[44] In some cases, this required us to "back calculate" the number of respondents from reported response rates and sample sizes (which, in the case of factorial designs, were not always reported for each subgroup), or to compute standard deviations from reported standard errors and sample sizes. We acknowledge that it is possible that rounding error resulting from these approximations may have caused us to conclude that the CIs for some RRs contained unity when this was not in fact the case, or vice versa. Because of the need for such "back calculations", to avoid spurious accuracy we have chosen to present findings as percentages, rather than as the number of events per experimental group (although insofar as was possible we have presented the total sample sizes for each group).

We initially considered following the model used in Cochrane-style reviews,[30] of attaching an explicit "level of evidence" to each conclusion and recommendation (the highest level equals evidence from multiple well-designed randomised trials). However, the balance of opinion within the research team was that this would be potentially misleading. Heterogeneity between studies makes comparisons across studies difficult. Findings from a randomised trial of reminders in a general population survey on a topic of low interest in the USA cannot readily be equated with a trial of reminders in a survey of patients' attitudes to healthcare (generally a high-interest topic) in the UK.

We therefore concluded that in our tables of results we should simply present the evidence without any attempts at meta-analysis, or even at detailed weighting. We have, however, included the "quality score" described above, and ordered the presented studies, from randomised controlled trials on health-related topics to non-random studies on non-health topics. Likewise, as indicated above, we separated recommendations derived from primary studies from those resulting from previous literature reviews, expert opinion or theories of respondent behaviour. Furthermore, we have highlighted those aspects of survey design and administration where there is a paucity of evidence from high-quality health-related studies. We suggest that individual readers then need to make a case-by-case judgement on whether the evidence presented is likely to be generalisable to their particular situation.

Limitations of the approach

Self-imposed limitations

The main limitation of our approach, as with any non-exhaustive review, is that, by focusing on a limited number of databases and confining our search to a specific period, we may have failed to identify a number of relevant studies. Concentrating on the MEDLINE and PsycLIT databases represents a bias in favour of published literature indexed on those databases. Articles in journals included only on other electronic databases, and those excluded from all the major databases, could not have been identified. The systematic exclusion of databases may lead to publication bias in favour of articles published in the more popular and renowned journals. Similarly, our exclusion of the "grey" literature means that we will have missed methodological studies of aspects of questionnaire design and administration published only in the form of internal reports from survey methodology organisations and so on. The major threat in excluding the grey literature and articles published in the more obscure journals not included in the MEDLINE and PsycLIT databases is the risk of exclusion of studies with negative findings or those showing lower effects than for studies published in more renowned journals. Such publication bias against negative trials has been recognised and is well documented in respect of clinical trials.[45,46] If such a bias were to have occurred in our identification of studies, we would be at risk of overestimating the effects of the interventions under investigation.

Moreover, even within our chosen databases, our decision to confine the search strategy to MeSH and thesaurus terms, rather than conducting a more detailed search of key terms within titles and abstracts, means that our identification process is only as good as the keywording employed by MEDLINE and PsycLIT. If references on relevant topics did not have one of our chosen keywords (surveys, questionnaires, interviews) added as a MeSH or thesaurus term, we would not have picked it up. Given the breadth of topics that we wished to cover, however, and the range of synonyms for each topic, we did not consider it feasible to develop a more focused search strategy.

Finally, our error in omitting some references identified through both the MEDLINE and PsycLIT databases (MEDLINE for 1987–1992; PsycLIT for 1979, 1991, 1993–1996; appendix 5) clearly represents a failure to locate a number of relevant trials, including some on health-related topics. The risk of systematically omitting papers identified in particular years is greatest when research is

incremental, with repeated studies on the same topic seeking to confirm or refute the findings of previous researchers, and to generalise those findings to new populations and settings. Focusing only on "early" or "late" studies may therefore be a source of bias, although the direction of any such bias cannot readily be predicted (initially positive findings could be refuted by later negative trials; conversely, early negative or null findings could be followed by later positive studies).

It is impossible to quantify the effects of our selection biases. Whether publication bias in favour of positive trials is as rife in the field of survey research as it is reported to be in respect of clinical trials[45,46] remains open to debate and investigation. We have certainly identified negative and null trials. The PsycLIT database in particular yielded a number of relevant articles from relatively obscure journals. It is our personal belief that, for some aspects of questionnaire design and administration, evidence from the studies we have identified appears to be conclusive. Moreover, findings from these empirical studies are supported by psychological and sociological theories of respondent behaviour, and accord with the accumulated experience and consensus views of survey methodology experts (the vast majority of whom are also engaged in experimental work on aspects of questionnaire design and survey administration). Where evidence is equivocal, the most usual and plausible explanation is that the effect of a particular approach or technique of data collection is context or topic specific. We believe that further evidence, from heterogeneous studies, is unlikely to resolve the current uncertainty.

Furthermore, in conducting our review, it appeared to us that research into questionnaire construction and survey administration is haphazard. We observed little sense of concerted effort to build on previous research or to generalise findings from one study to other settings, populations or modes of administration. In contrast to clinical research, a systematic and incremental approach appeared to be the exception rather than the rule. For this reason, we consider that the exclusion of complete years of studies identified through our chosen databases, although a systematic effect in itself, is unlikely to have resulted in systematic bias with respect to the evidence. In other words, there is no good reason to assume that studies reported in those omitted years would all report findings in the same direction or would be on topics not researched elsewhere.

Externally imposed limitations

External limitations – in particular, the ways in which studies were designed and reported – restricted the interpretation, comparison and generalisation of the evidence obtained.

The lack of information in the identified articles on certain aspects of study methodology imposed problems in assessing the quality of these studies and in deriving and interpreting quality scores. In particular, it was rare to find details of power calculations in determining sample size; in most cases, sample size appears to have been determined by the requirements of the "parent study" rather than by the power to detect differences with respect to the factors manipulated. Almost all identified studies "lost a point" for this omission. Studies were also downgraded for not holding all other factors constant, but in some cases there were good reasons for this, for example, the impossibility of deriving "long" questionnaires that contained only factual questions.[47] In general, study quality (at least as assessed according to the criteria we chose) was fairly uniform, precluding the ranking of studies by quality score.

A major problem in the comparison and interpretation of results was posed by heterogeneity between studies. Comparisons of, for example, response rates across studies in which the same aspect of survey administration (e.g. incentives) was manipulated were hampered by the lack of comparability with respect to other aspects of questionnaire design and administration (e.g. length of questionnaire, number of reminders sent) and by differences in study populations. This heterogeneity precluded meta-analyses.

A related issue was that few researchers replicated exactly in a new setting the work of previous researchers, to see whether the findings were generalisable or if they were context or mode specific (e.g. whether the effects of personalisation were the same in surveys of general and special populations). The exception to this was in some of the studies regarding question wording, where researchers tested whether the question sequencing effects observed in interview surveys also occurred with self-completion questionnaires. This apparent desire for "uniqueness" ran counter to the recommendations for future research made in many articles, which explicitly recommended that the generalisability of the findings should be confirmed by further research in different settings.

In designing trials of aspects of survey methodology, many researchers did not appear to take into

account adequately the complexity of the survey process, in particular the way in which a number of factors interact to influence respondent behaviour.[1] This lack of consideration of the likely interactions led, in some cases, to inappropriate and unrealistic manipulations (e.g. a highly personalised letter with an assurance of anonymity).

We identified ethical issues in applying some of the techniques advocated by other survey researchers. For example, Hornik[48] showed that telling respondents that a questionnaire would take less time to complete than was actually required led to an improved response rate; however, deliberately deceiving respondents in this way is likely to be viewed as unethical, at least in health surveys. Similarly, although the provision of incentives has been shown to be a powerful means of stimulating response, paying survey respondents is generally frowned upon in health-related research.[3]

We also identified practical barriers in applying some of the techniques recommended. For example, Dillman[1] strongly advocated that covering letters should be individually signed. In a large survey this would almost certainly be unfeasible. We noted that most researchers were concerned with the effectiveness of the techniques they examined, rather than with their efficiency. Emphasis was placed on response rates rather than on the cost per response; few tested whether the additional costs of resource-intensive methods (e.g. multiple follow-ups or incentives) outweighed the marginal benefits in terms of increased response. Indeed, only in a minority of identified studies did the authors consider and report resource implications.

This lack of consideration of cost outcomes reflected the fact that many researchers considered only single measures of success in evaluating the impact of the factors they manipulated. Overall response rate was often the key outcome. Other important indicators of the quality of the data yielded, such as item response rates or the validity of the information provided, were frequently ignored. However, increases in questionnaire response rates may potentially be at the expense of quality of response.

Finally, the technologies available to survey researchers are constantly evolving and the world in which surveys are carried out is constantly changing. In the future it is likely that computer-assisted interviewing and computer-scannable (optical mark reading (OMR) and optical character recognition (OCR) technology; internet delivery) questionnaires will increasingly become the norm. Attitudes towards surveys among the public and professionals are also likely to change; already, a downward trend in response rates for health professionals has been observed.[49] Recommendations made on the basis of evidence and expert opinion in the 1990s may no longer hold good in the twenty-first century.

Implications for users of this review

The limitations imposed upon us, as well as those we imposed upon ourselves, limit the extent to which universal recommendations of best practice in designing and using questionnaires with patients and staff can be made. In particular, caution needs to be exerted in overgeneralising from non-health-related research to health surveys, from one culture to another, across populations, and in extrapolating the findings from one mode of administration to another.

Theories of respondent behaviour and the findings from this review indicate that no single method of questionnaire administration consistently outperforms all others. Rather, the choice of method will depend on maintaining a balance between the volume and quality of data required and the resources available to complete the survey.[24,25] Similarly, the most appropriate wording and sequencing of questions and of response categories depends on the study population, the survey topic, the specific information to be gathered, and the mode of administration. There is no single method of enhancing response rates that is applicable in all settings. Instead, the choice of techniques should be informed by consideration of the likely barriers and motivational factors for each particular survey topic and study population.

In designing and administering a survey, each researcher needs to consider the particular circumstances of that survey and to address the following questions.

- Who is being surveyed?
- Where?
- When?
- What information needs to be collected?
- In what detail?
- What is the desired accuracy?
- What level of accuracy is reasonably attainable?
- What resources (time, money, personnel, skills) are available?

As suggested above, a further review of the relevant literature will often be required to inform the decision on if a questionnaire survey is indeed the most appropriate method of data collection for that specific population, setting and topic.

Judgements about whether the recommendations of the experts are feasible and if the evidence from existing studies is applicable in a particular circumstance need to be made on a case-by-case basis.

This review should be reviewed as a **guide** to best practice, not a **definition** of best practice. It is a decision aid, not a substitute for critical appraisal of the options available.

Chapter 3

Methods of survey administration

Introduction

One of the first decisions to be made in designing and conducting a survey is that regarding the mode of administration. Essentially the choice is between interviewer administration (either face-to-face or by telephone) or self-completion by the respondent (with delivery of the questionnaire either by post or to a "captive audience", such as employees at their work place or patients attending a clinic). Both face-to-face and telephone interviews can be computer assisted (the terms computer-assisted personal interviewing (CAPI) and computer-assisted telephone interviewing (CATI) are widely used). There is also growing interest in the computerised administration of self-completion questionnaires (computer-assisted self-administration (CASI)). When paper-based questionnaires or interview schedules are used, data entry to computer can be facilitated by the use of OMR and OCR, otherwise known as scannable questionnaires.

Each mode of administration has its advantages and disadvantages.[1,50] These are summarised in *Table 1* and discussed in greater detail below.

TABLE 1 *Advantages and disadvantages of modes of questionnaire administration (adapted from de Vaus,[50] after Dillman[1])*

	Face-to-face interviews	**Telephone interviews**	**Postal questionnaires**
Response rates:			
General population samples	Usually best	Usually lower than face-to-face	Poor to good
Special population samples	Usually good	Satisfactory to best	Satisfactory to good
Representative samples:			
Avoidance of refusal bias	Depends on good interviewer technique	Depends on good interviewer technique	Poor
Control over who completes the questionnaire	Good	Moderate	Poor to good
Gaining access to a named selected person	Good	Good for those with telephones	Poor to good
Locating the named selected person	Good	Good	Good
Ability to handle:			
Long questionnaires	Good	Moderate	Satisfactory to poor
Complex questions	Good	Moderate	Moderate to poor
Boring questions	Good	Moderate	Poor
Item non-response	Good	Good	Moderate
Filter questions	Good	Good	Moderate to poor
Question sequence control	Good	Good	Poor
Open-ended questions	Good	Good	Poor
Quality of answers:			
Minimise social desirability responses	Poor	Moderate	Satisfactory
Ability to avoid distortion due to:			
Interviewer's characteristics	Poor	Moderate	Good
Interviewer's opinions	Moderate	Moderate	Good
Influence of other people	Moderate	Good	Poor
Allows opportunities to consult	Moderate	Poor	Good
Implementing the survey:			
Ease of finding suitable staff	Poor	Moderate	Good
Speed	Poor	Good	Poor
Cost	Poor	Moderate	Good

Face-to-face interviews

In interviewer-administered surveys, the interviewer asks respondents questions based on the assumption of "equivalence of stimulus".[13] In other words, as far as possible, each respondent is asked the same questions, with the same meaning, using identical wording and sequence of words. The aim is to eliminate any bias that may be caused by differential stimulus; for example, if the wording of a question varied from respondent to respondent, any observed differences between respondents could be due to this variation in stimulus rather than to true inter-respondent differences in attitudes.

Advantages

In an interviewer-administered survey, the burden of recording the responses lies with the interviewer rather than with the respondent. Because of this, and because the interviewer can probe and prompt for further details, interviewer administration facilitates the collection of larger amounts of information, and of more detailed and complex data. Questionnaires administered by interviewers facilitate the use of open-ended questions, or open-ended probes, where the interviewer can record verbatim the answers given by respondents. This may generate richer and more spontaneous information than would be possible by using self-completion questionnaires. Although open-ended questions can be used in self-completion questionnaires, responses are typically less detailed because the burden of recording the response falls on the respondent.

Because of the interpersonal interaction, response rates to interview surveys are typically higher than for postal surveys; non-response bias is therefore likely to be less of a problem. Furthermore, interview surveys may reduce sample composition bias by ensuring that information is actually obtained from the target respondent. Self-completion questionnaires, on the other hand, rely on self-selected samples (i.e. those who complete and return the questionnaires), with findings not necessarily being generalisable to the underlying population. Moreover, with self-completion questionnaires (especially those sent through the post), it is impossible to be sure that the questionnaire has in fact been completed by the target respondent and not by another member of the household.

Interviewers may be able to provide respondents with a more convincing explanation of the purpose of the study than would be possible in a covering letter for a self-completion questionnaire. They can use their powers of persuasion, thereby stimulating participation rates.[51] According to Oppenheim,[13] interviewers should also be able to engage respondents' interest and attention, thus leaving them "feeling that something pleasant, interesting and worth while" has been accomplished.

Interviews enable researchers to reach less well-educated respondents and to obtain answers from people with reading and writing difficulties. They are also appropriate for gathering information from people whose language has no written representation. Surveys of ethnic communities, or of users of sign language, can be facilitated by the employment of interviewers who are competent in the languages of the target population; bias in interpreting and relaying questions and responses is a risk if a translator has to be used as an intermediary.

Interviewers can also enhance the quality of data collected by offering clarification and explanation of any problems arising in the course of the interview, reducing misunderstandings (although this may introduce interviewer variance) and ensuring that questions are answered in the correct sequence. The use of filter questions and complex skip instructions are facilitated by this mode of administration, and visual aids (e.g. prompt cards for multiple response questions) may be employed. Finally, interviews enable on-the-spot verification of issues that are relevant to the survey (e.g. visual assessment of whether or not the respondent is obese in a survey on health and life-style). This ability to validate or verify respondents' answers may reduce the threat of response bias; respondents may, for example, be less likely to report themselves as being younger than they actually are.

Disadvantages

A downside of interviewer administration is that this method gives researchers the opportunity to go down more complex, time-consuming and costly avenues. The ability to collect more data, and data of greater complexity, can lead to the temptation to gather more information than is actually needed for the study. Interviews are more expensive than postal surveys; the additional expense derives mainly from the costs of training and then paying interviewers (whether on a per-interview or per-hour basis) and of travel costs. Interviewer-administered surveys typically contain more open-ended questions than do self-completion questionnaires; coding these can be a time-consuming and costly procedure. It follows, then, that studies using interviewer-administered questionnaires may take longer to produce results

than postal surveys. Balanced against this, however, researchers using interviewer-administered surveys do not have to wait for questionnaires to be returned through the post.

Another significant disadvantage of interview surveys is that interviewers can introduce errors in both a random manner (variance) and a systematic way (bias).[13] The former is likely to be due to interviewer inaccuracy, for example, random errors in recording answers or altering the wording of a question by mistake. Examples of the latter are: selective, rather than verbatim, recording of participants' responses; "differential probing"[5] – differences between interviewers in the extent to which they probe for a substantive response or accept "don't know" answers; and consistent rewording of questions. Systematic bias can occur even when there is only one interviewer if the interviewer does not record accurately the respondent's answers verbatim, but instead is consistently selective in what is recorded. It is an even greater problem when a survey requires multiple interviewers; observed differences across respondents may be an artefact of the way in which different interviewers have posed questions or recorded answers, rather than an indication of true underlying differences between participants.

"Recency" effects, whereby respondents choose response categories towards the end of multiple choice lists are more likely to occur in interview surveys[52] because of fatigue effects and memory effects (the respondent is more likely to remember the last options read out by the interviewer). The social interaction between the respondent and the interviewer can also be a source of response bias, particularly social acquiescence bias whereby respondents give the answer that is "socially desirable" or shows them in a good light, rather than reporting their true behaviour or attitudes.

Personal characteristics of the interviewer, such as age, gender, social class, race or level of experience and training, may also affect both response rates and the nature of the responses given. However, studies of interviewer variance are bedevilled by the fact that different interviewers obtain different response rates, so it is hard to disentangle non-response bias from response bias. The literature shows no consistent trends with respect to the impact of interviewer characteristics on either response rates or the type of responses given. It is likely that the impact of interviewer characteristics may be modified, or even confounded, by study population, survey topic and by how sensitive or embarrassing the questions are. For example, although it is often postulated that a mismatch in race or ethnicity between interviewer and respondent can affect the responses obtained, the consensus from a number of studies[53–59] is that such effects are likely to be confined to sensitive (i.e. race-related) questions.

Telephone interviews

Advantages

Telephone interviews are seen as a means of maximising the advantages of using interviewers while minimising the disadvantages. By eliminating travel costs and time, telephone interviews can be a low-cost and speedy method of data collection.[13,60] However, costs increase with the number of attempts made to contact people who are not available at the first call and with the number of long-distance calls required.

Another advantage of telephone interviews is that stricter control and closer supervision of interviewers is possible by comparison with face-to-face interviews[51] because interviewers can be monitored by supervisors listening in, with the consent of the respondent and the interviewer. This greater control can reduce inter-interviewer variability.

Telephone surveys generally have a reasonably high response rate. However, because of the lack of direct contact between interviewer and respondent, response rates may not be as high as with face-to-face interviews. Non-response is typically 5–10% higher than in comparable face-to-face surveys but, as with any survey method, response is higher when the topic is of direct interest to participants.[51] For long interviews or non-salient topics, however, completion rates for telephone interviews may be considerably lower than for face-to-face interviews, since it may be socially more acceptable to refuse or terminate prematurely an interaction over the telephone.

Telephone interviews are said to be suitable for all but the most complex questions, but ideally there is a need to avoid questions with a large number of possible responses because respondents will not be able to keep the information in their heads long enough to answer reliably.[50] They are also held to reduce resistance to sensitive questions and decrease the tendency to give socially acceptable responses because of respondents' relative anonymity.

Interviewer effects are lower. Visible characteristics of interviewers cannot influence answers,[60] but accent may have an effect on comprehension and

the technique may be unsuitable for surveys of special populations (e.g. those who are hearing impaired) unless special equipment is employed. Telephone interviews also allow survey researchers to access geographical areas where interviewers' safety may be threatened, for example, inner cities.[51]

Disadvantages

The obvious disadvantage of telephone interviewing is the problem of sample composition bias and the effect on the generalisability of findings. Those with lower incomes, young people, those who have recently moved house, and ethnic minorities are less likely to have a telephone.[13] This could lead to significant bias if the sector of the population that is inaccessible by telephone is the subject of the research. However, the proportion of households (and the attendant bias against younger people and those living in urban areas) who are ex-directory (approximately 37% – data supplied by British Telecom) is now potentially more of a problem to survey researchers than those not having a telephone (now only about 4% of households, according to the 1998 General Household Survey). This issue is relevant only for those who select their sample from the telephone book; for some surveys (e.g. of ex-patients), telephone numbers are known already. Where telephone numbers are unknown or unlisted, random digit dialling techniques may be employed, but this approach is likely to result in a high proportion of unconnected numbers and a significant rate of ineligible contacts. The likely impact on response rates of the increasing use of answering machines and call-screening services, and of mobile phone ownership, remains to be determined.[61,62]

Telephone interviewing may also be problematic with people who are hard of hearing, elderly or from minority ethnic groups (unless interviewers speaking the same language are used), but these problems may be no greater than with face-to-face interviews.

In telephone interviews, there is no scope for using visual aids (e.g. prompt cards). Recency effects may therefore be exaggerated and the use of long lists of multiple responses or complex response formats is precluded. Furthermore, there are no visual cues giving the interviewer information about participants' reactions to questions or their ability to communicate non-verbally (e.g. smiling, eye contact etc.).

Self-completion questionnaires

Delivery and return through the mail (i.e. postal surveys) is the most common mode of administration for self-completion questionnaires.

However, supervised self-completion ("captive audience") surveys are also used. Here, respondents complete questionnaires in the presence of a researcher, who is available to provide some assistance or explanation and who may also check questionnaires for completeness of response. This technique can be used for both individuals and for groups (e.g. students in a classroom, employees in a workplace setting).

Advantages

The main advantage of self-completion questionnaires is their low cost compared with other methods. In interviewer-administered surveys, geographical clustering (a two-stage sampling process in which geographical units, such as electoral wards, are initially sampled and then a sample of population units, such as households, are selected within each sampled geographical area) may be necessary to avoid high travel costs; this may result in less precision for a given sample size. In contrast, postal surveys can cover widely dispersed populations without increasing study costs. Similarly, in captive audience self-completion surveys, data can be collected simultaneously from a large number of respondents.

Bourque and Fielder[40] gave two further sample-related advantages. Postal surveys allow researchers to study larger groups and they provide wider coverage within a given study population, particularly among those who are reluctant to be interviewed in person or on the telephone. Postal surveys may also be quicker than interviewer-administered surveys, although time must be allowed for late returns and follow-up attempts. Postal surveys are easier to implement and require fewer personnel (in comparison with all forms of interview survey) and minimal equipment (compared with telephone interviews).

Unlike other methods of data collection (in particular face-to-face interviews), it can generally be assumed that all potential participants receive the mailed questionnaire at approximately the same time, therefore "context" or "history" effects that may influence their experiences, opinions or attitudes are minimised for the total sample. (This would not be true, however, if a high proportion of the sample were away from home when the questionnaire arrived.)

No interviewer is involved in self-completion questionnaires, so they avoid the potential for interviewer bias as described above. They may also be more appropriate if information is required about several members of a household or if an

answer requires the consultation of documents (e.g. "When was your last hospital appointment?"). Participants may respond more truthfully to sensitive questions by using this approach, and may make more critical or less socially acceptable responses than when face-to-face with an interviewer.

Finally, postal surveys avoid the problem of respondents being unavailable when the interviewer calls.

Disadvantages

Bourque and Fielder[40] divide the disadvantages into three sections: "sample-related", "questionnaire construction" and "administration".

To consider sample-related factors, selecting a sample representative of the population of interest is dependent on obtaining a complete and accurate list of the population to act as a sampling frame; however, these lists may be unavailable, incomplete or inaccurate. (Of course, this is also a problem in other methods of survey administration.) Secondly, the biggest disadvantage of postal surveys is their typically lower response rates. Although these can be increased, for example, by the use of follow-up mailings and incentives (see chapter 6), response rates are generally held to be lower than for face-to-face or telephone interviews. One reason for this is that potential participants who have literacy problems, or visual or motor impairments, are unable to respond, as are some of those who speak a different language from that used in the questionnaire. It is also easy to forget, ignore or mislay a mailed questionnaire. An important point is that with postal surveys there is a greater likelihood that respondents will differ significantly from non-respondents (the most obvious bias being in favour of better-educated and more literate individuals), so estimates based on achieved responses may be biased.[5] Although non-response bias is an issue in all approaches, the problem is greater in postal surveys, where relatively lower response rates are common.

Turning to questionnaire construction, postal surveys are best suited to clear, non-complex research topics that are capable of being explained in a few paragraphs.[5] This has implications for questionnaire length and format; it is generally suggested that self-completion questionnaires should be shorter than interviewer-administered questionnaires and should contain mostly closed-ended questions without branches and skips. The questionnaire must also "stand alone"[40] and should be as easy as possible to complete without assistance. There is no opportunity to probe beyond the answer given, clarify ambiguities or overcome unwillingness to answer a particular question. Thus, according to Moser and Kalton, postal surveys are an "inflexible method" (p. 260).[5] The researcher has no control over the order in which respondents answer questions, so postal surveys may be less appropriate when answers to one set of questions could bias or otherwise influence answers to another section of the questionnaire. "Primacy" effects, whereby respondents select the first response that seems applicable, without considering the full range of alternatives, may be more common in self-completion questionnaires.[52]

Thirdly, there are disadvantages in relation to administration. With postal surveys in particular, the researcher has no control over who completes the questionnaire, or whether they consult with others. Although postal surveys may be quicker to perform than interviewer-administered ones, the use of follow-up mailings, and perhaps telephone reminders, to boost response rates means that the data collection period could extend over several months. In order to achieve an acceptable response, the survey budget may need to allow for at least one reminder, with an additional copy of the questionnaire (see chapter 6).

Finally, in postal surveys there is no opportunity for supplementing answers with observational data.

Computer-assisted approaches

The use of CAPI and CATI may help to minimise the random and systematic interviewer errors outlined above. In this approach, the interviewer uses a computer terminal rather than a paper questionnaire and keys in answers to questions as they appear on the screen. Tailoring of the questionnaire to the individual respondent, in particular the implementation of skipping and branching, is facilitated. This use of technology may prevent routine errors and omissions in asking questions, recording responses and following complex skip patterns. Moreover, results and response rates are available quickly because the intermediate steps of data editing and entry are eliminated. However, both CAPI and CATI at present require high levels of investment in purchasing hardware and programming computers; these set-up costs may be prohibitive for many survey researchers.

There is also growing interest in, and use of, computer technology for the delivery of self-completion questionnaires. One alternative to

manual data entry from paper-based questionnaires is electronic scanning. This must be planned into the document design and printing, and requires appropriate computer hardware and software. Respondents must be told how to record responses (e.g. tick a box using a black pen). Scanning is less restrictive on the positioning of questions on the page than conventional manual keying designs. It is particularly appropriate for closed questions, where a box or bubble can be ticked or filled in and OCR software can be used. However, it is less effective for "write-in" answers; special arrangements (OMR) need to be made to capture numerical and alphabetical characters. Questionnaires may require to be dismantled for scanning and some completed questionnaires will fail to scan correctly.

Analogous to CAPI and CATI is the use of a CASI. This may require the respondent to use a computer keyboard for data entry, or may employ touch-screen or light-pen technology. There is also increasing interest in the delivery of questionnaires over the Web or by e-mail.[63]

Identified studies

In total, 17 randomised controlled trials or non-random concurrent controlled studies[64–80] that compared some combination of face-to-face interviews, telephone interviews and self-completion questionnaires were identified. None reported that sample sizes were based on a power calculation. Quality scores for several studies were also affected by the fact that factors other than mode of administration were not held constant. However, the authors of the original articles generally argued that there were practical reasons for the lack of consistency of treatment (e.g. different strategies with respect to the number and timing of contacts being appropriate for different modes of administration).

Criteria for assessing the relative performance of different modes of survey administration included:

- Instrument response rates: the percentage of the target sample who agreed to participate (for interview surveys) or returned a questionnaire (for self-completion surveys). In some cases, response rates were adjusted to take account of non-contact, ineligibility or number of questions completed.
- Non-response bias: the extent to which respondents and non-respondents differed with respect to important variables that were likely to influence substantive findings. In some cases, comparisons were in terms of achieved sample composition by the different means of administration; this gives an estimate of relative, but not absolute, non-response bias.
- Item non-response rates: the number of items omitted or with a "don't know" response given. This provides a proxy for how understandable and acceptable the questions were.
- Quality of data provided: the volume and nature of the information obtained and the extent of response bias (bearing in mind that rarely was there a "gold standard" by which data validity could be assessed).
- Resource use: both financial and non-monetary resources (e.g. time).

Self-completion questionnaires versus telephone interviews

Four studies were identified in which telephone interviews were compared with postal questionnaires.[64–67] Three were randomised controlled trials and one was a non-random concurrent controlled study (*Table 2*; see p. 32). All were on health-related topics and all were carried out in community settings in North America.

Instrument response rates

All four identified studies examined the effect of mode of administration on instrument response rates, with two also reporting non-contact rates. Pederson and colleagues[67] showed a significantly higher non-contact rate for the telephone approach but found that refusal rates were significantly lower for this method of administration. In contrast, Hinkle and King[64] reported a lower non-contact rate for the telephone survey, but the difference was not statistically significant. In three studies,[64,65,67] higher questionnaire completion rates were found for telephone interviews. In the studies by Talley and colleagues[65] and by Hinkle and King[64] the differences were statistically significant. The differences found by Pederson and colleagues[67] were not statistically significant, but their sample was small and therefore lacked power. McHorney and colleagues,[66] on the other hand, found significantly higher response rates from their postal survey (79% versus 69%). However, in this study, not all questionnaires were completed by the originally assigned mode of administration; 65% of those originally assigned to postal administration actually returned the completed questionnaire by mail, while the remaining 14% had to be interviewed by telephone; among those originally assigned to telephone administration, 65% completed the telephone interview, and 4% completed a postal version. When those completing the survey by a

mode other than that originally assigned were excluded, the difference in response rates between postal and telephone administration was not statistically significant.

Non-response bias

McHorney and colleagues[66] reported that non-response bias was found in both modes, with non-respondents being less well educated and having a lower income, and being less likely to be in employment. Compared with respondents for the same mode of administration, postal non-respondents were more likely to be younger, non-white and either single, separated or divorced, while telephone non-respondents were more likely to be male. However, only with respect to age was the nature of non-response bias different between the two modes of administration; respondents to the postal questionnaire were older than non-responders.

Hinkle and King[64] compared the socio-economic status of their achieved samples with census data on income level. They found that, although all methods over-represented those with higher incomes and of higher socio-economic status, the achieved sample for the postal survey was particularly biased in favour of this group, especially for those who had utilised mental health services.

Item non-response rates

The study by McHorney and colleagues[66] was the only one to examine item non-response rates, one indicator of the quality of response. The mean number of missing responses was significantly lower for telephone administration.

Nature and quality of data

Talley and colleagues,[65] McHorney and colleagues,[66] and Hinkle and King[64] looked at the content of information provided by both modes of administration. Talley and colleagues[65] found that telephone respondents rated specific psychological needs as of more importance to them than did respondents to the postal survey; they suggested that this could be indicative of "social desirability" bias. However, items given the most importance by respondents to the telephone survey were also accorded high importance by respondents to the postal survey. McHorney and colleagues[66] found that their postal survey yielded less favourable health ratings than the telephone survey. They concluded that postal surveys offer more anonymity for reporting sensitive and personal information. Hinkle and King[64] reported that, in the mail survey, respondents of higher socio-economic status tended to give more neutral and negative responses about mental health services, especially if they had received help from those services.

Pederson and colleagues[67] looked at the truthfulness of reports provided by both modes of administration. In a subsample of respondents, reported smoking status was validated by a test of salivary thiocyanate; no lying was detected. In this study there were also no detected consistent differences between the two modes of administration in respect of making "socially desirable" responses.

Resource use

Three studies compared the financial cost of each method, again with equivocal findings: two found the postal survey more expensive,[64,67] while McHorney and colleagues[66] found costs for the telephone survey to be 77% higher than for the postal survey.

Summary of findings

From these studies there is little consensus about the relative benefits of telephone and postal surveys on the parameters of instrument completion rates, non-response bias, quality of response, anonymity or cost. The only parameter showing agreement in three out of four studies was response rate, which was generally higher for telephone interviews.

Self-completion questionnaires versus face-to-face interviews

Seven studies in which face-to-face interviews and self-completion questionnaires were compared were identified;[64,68–73] six were randomised controlled trials (*Table 3*; see p. 34). Boekeloo and colleagues[73] compared an audio self-administered questionnaire (delivered by a cassette player and headset) with a written self-completion questionnaire; the questionnaires were then followed by face-to-face interviews. Liefeld[72] carried out a three-way comparison of paper-based self-completion, computer-assisted self-completion, and face-to-face interviews. All but two[69,72] of the identified studies were on health-related topics.

Instrument response rates

Two studies measured the effect of mode of administration on instrument response rates;[64,71] in both, response rates were significantly higher for face-to-face interviews. Hinkle and King[64] also reported a significant difference in non-contact rates between the two modes, with a lower rate observed for face-to-face interviews.

Non-response bias

Non-response bias was examined by Cartwright,[71] who found that Asian mothers were under-represented among those responding to a postal survey about women's experiences of maternity services in respect of a recent birth. In the achieved

postal survey sample, only 2% of the babies' parents were both Asian-born, compared with 7% in the achieved sample for an interview survey, and 7% in the target sample.

Item non-response rates

Three studies measured the effect of mode of administration on item non-response rates; the findings were inconsistent. Newton and colleagues[68] found that respondents were more likely to give no answer at all or to say "don't know" in an interview than in a self-completion questionnaire or card sort. For 84 out of the 202 items in the questionnaire, there was a significant effect of mode of administration on item non-response rates; for 81% of these items, the highest rate of missing responses was for interviewer administration. In Cartwright's study,[71] item non-response rates were generally low, but they were nonetheless greater for postal administration (mean missing items 1.9% versus 0.6% for interview). Boekeloo and colleagues[73] found that the mean level of item non-response was highest for written self-administration.

Nature and quality of data

Six studies[68–73] examined responses to sensitive questions and the possibility of social desirability biases, with equivocal findings.

Newton and colleagues[68] found that items that were high in social desirability were no more subject to the effects of mode of administration than items low in social desirability. However, respondents were more willing to endorse negative items (e.g. "I dislike my job") or reject positive items (e.g. "I like my job") in interviews, compared with self-completion questionnaires or card sorts.

Oei and Zwart's study of life events[70] presented conflicting evidence: the mean number of events reported was higher in the more anonymous self-completion questionnaires but the types of events reported varied according to the mode of administration. The frequency of reporting death was higher in interviews, while life events relating to working conditions, education and training, illness (self) and marital problems were reported more often in self-completion questionnaires.

Cartwright[71] found no major differences in the nature of response; in particular, replies to painful or delicate subjects were similar in both groups. However, she concluded that there was some support for the view that criticisms would be reported more in an interview rather than in self-completion mode, especially when open-ended questions are posed.

Boekeloo and co-workers[73] found that more HIV risk factors were identified in self-completion questionnaires: when paper-based self-completion questionnaires were compared with interviews, the rate of reporting risky behaviour was significantly greater for four out of 16 items; the rate of reporting risky behaviour in response to an audio-administered self-completion questionnaire was significantly greater than for interview for six out of the 16 items.

Liefeld[72] reported few differences in response patterns by mode of administration for factual questions, especially those requiring a dichotomous – "yes" or "no" – answer. For multiple-response questions testing knowledge of shopping facilities, respondents to computer-assisted self-completion questionnaires gave more incorrect answers than did those completing paper-based questionnaires or participating in a face-to-face interview.

Nederhof[69] found that a higher number of altruistic and socially desirable responses were given in face-to-face interviews compared with postal questionnaires.

Resource use

Hinkle and King[64] found that cost per completed questionnaire was higher for postal administration.

Summary of findings

Overall, there is no good evidence to support the view that postal survey participants respond more truthfully to sensitive issues and make more critical or less socially acceptable responses than when face-to-face with an interviewer. In line with what is commonly asserted, the limited evidence from the identified studies suggests that response rates are higher for face-to-face interviews.

Telephone versus face-to-face interviews

Five studies that compared telephone and face-to-face interviews were identified.[64,74–77] Three were randomised controlled trials and two were non-random concurrent controlled studies (*Table 4*; see p. 36); all but one[75] were on health-related topics

Instrument response rates

All five studies reported the effect of mode of administration on response rates, with mixed results. Jordon and colleagues[74] found that a significantly lower proportion of those sampled for face-to-face interview were ineligible for inclusion; this was probably due to differences in the methods by which the two groups were selected (the sample for face-to-face interviews was drawn from a sampling frame of computer-readable addresses selected on an area probability

basis, while the sample for telephone interviews was drawn by adding four random digits to a sample of telephone numbers selected from reverse telephone directories). These researchers also reported a significantly higher response rate in the face-to-face interview group.[74] Aneshensel and colleagues[76] found a slightly higher response in the telephone interview group (but this was not significant). Hinkle and King[64] reported no difference between the two modes with respect to non-contact rates, but, once contact had been made, questionnaire completion rates were higher among those participating in a face-to-face interview. Fenig and colleagues,[77] in contrast, reported a higher response rate for their telephone interviews; however, in this study, the face-to-face interviews were always carried out after the telephone interviews, so fatigue or disaffection may have accounted for this discrepancy. Finally, Quinn and co-workers[75] found there was no significant difference in overall response rates or when rates for male and female interviewers were compared; however, in this latter study, refusal rates were significantly (just) higher for telephone interviews.

Non-response bias

Quinn and colleagues[75] found differential sample composition with respect to gender between the two modes of administration. The target respon-dent was "responsible adult in household"; 70% of respondents to the telephone interview were female, compared with 55% for face-to-face administration. However, they found no further evidence of sample composition bias.

Item non-response rates

Jordon and colleagues[74] found that respondents to the face-to-face interview had significantly less missing data on family income. The authors reported no significant difference in the number of responses to open-ended questions between the two modes of administration. However, there were significantly more (and contradictory) responses to checklist questions under telephone administration, which the authors attributed to the need for the interviewer to read each response option individually and to elicit a "yes" or "no" response in the context of a telephone interview. Quinn and colleagues[75] observed slightly higher item non-response rates in telehone interviews.

Nature and quality of data

Fenig and colleagues[77] found a significantly higher mean score on the "rate of demoralisation" scale (indicating greater demoralisation) for the telephone mode. They concluded that telephone interviews are appropriate even in highly sensitive populations.

In contrast, Jordon and co-workers[74] reported that telephone interviews generated more "acquiescence" (a variable obtained by scoring all items in the "agree" direction, regardless of item content) and "evasiveness" (a variable computed by counting all "don't know" and "no answer" responses); it also resulted in contradictory answers to checklist multiple response questions. They concluded that this mode gave inferior-quality data.

In Aneshensel and colleagues' survey[76] of community health status, a significantly greater number of interview respondents reported restricted activity days (due to disability) compared with telephone respondents.

Resource use

Hinkle and King[64] noted that telephone interviews were cheaper than face-to-face interviews. Quinn and colleagues[75] compared interview completion times, finding that telephone interviews were quicker than face-to-face interviews.

Summary of findings

In summary, these studies do not provide a consistent picture of whether face-to-face interviews or telephone interviews are superior on the parameters of instrument response rate, eliciting sensitive information and item non-response.

Computer-assisted versus paper-based self-completion questionnaires

Four studies comparing computer-assisted and conventional questionnaires were found.[72,78–80] All were randomised controlled trials on non-health topics (*Table 5*; see p. 39).

Instrument response rates

Two studies measured the effect of mode of administration on response rates. Higgins and colleagues[79] found that there was no significant difference in crude response rates between interviewees using diskettes programmed with a questionnaire (DISKQ) and those completing conventional paper-based questionnaires. However, on adjusting for "ability to respond" (i.e. possession of an appropriate computer), response rates for the computer-assisted approach were significantly higher. Allen[78] reported that the response rate in the computerised questionnaire group was significantly lower than for a paper-based (but computer-scannable) questionnaire; however response rates were generally low in both groups (49% versus 29%).

Non-response bias

Allen[78] found no significant differences in respect of achieved sample composition between those

completing paper-based and computer-assisted questionnaires, a finding echoed by Higgins and colleagues.[79]

Item non-response rates

Higgins and colleagues[79] reported higher rates of non-response to a potentially sensitive question on income for those completing the DISKQ questionnaire, but this differences did not reach statistical significance.

Nature and quality of data

Response quality was assessed by Higgins and colleagues[79] and by Liefeld:[72] the former found the quality of responses (in terms of number of ideas generated and the verbosity of responses) to open questions was significantly better for the DISKQ method; the latter reported that respondents to computer-assisted questionnaires picked more incorrect answers for multiple response questions that tested knowledge.

Allen[78] found that respondents to the computerised version of the questionnaire produced a higher standard deviation and used a significantly wider range on rating scales than the paper-based self-completion group. He suggested that computer respondents "open up" more and tend to use slightly more extreme scale values. This may be due to a tendency to view the computer as more private, enabling more honest responses. However, in attempting to validate responses by reference to university records (assuming that these records themselves were correct), he did not demonstrate any consistent advantage of one mode of administration over another.

In contrast to Allen,[78] Helgeson and Ursic[80] found no significant difference in overall mean ratings or variances of ratings. Similarly, in Liefeld's study,[72] there were few response differences for factual (e.g. "yes/no") questions. The authors concluded that researchers can confidently compare results for factual questions from computer-assisted interviews, face-to-face interviews and self-completion questionnaire surveys.

Higgins and colleagues[79] examined the appropriateness of the two modes for posing sensitive questions; responses did not differ significantly. They concluded that bias with respect to mode of administration was not apparent in responses to sensitive questions, but that more research was needed in this area.

Resource use

Two studies[78,79] also concerned practical issues. Significantly quicker responses (i.e. time to return a completed questionnaire) were obtained using DISKQ,[79] while Allen[78] conducted a respondent evaluation and found that more computer participants would recommend the survey to a friend, although they believed the survey to be "too short". The group completing the paper-based, computer-scannable questionnaire perceived the length to be "about right".

Summary of findings

Again, there is conflicting evidence concerning the benefits of either mode on the parameters of response rate and response quality; only one study examined response speed and appropriateness of the modes for sensitive questions. None of these studies was in the health field, so findings should be extrapolated with caution.

Additional study identified

One study comparing CATI with conventional telephone interviewing was identified.[81] Response rates were significantly higher for the non-CATI group (82% versus 79%; $p < 0.05$; insufficient data were reported to calculate the RR and associated CI) and the time taken to complete a CATI interview was longer (52 versus 46 minutes; significance not reported). On most of the criteria measured (opinions of interviewers and respondents and most health statistics), only small differences were found between CATI and non-CATI approaches, but there was some evidence that interviewer variability was lower and there were fewer skip error problems in CATI (although the latter was not significant).

Conclusions

Self-completion questionnaires versus telephone interviews

- Telephone interviews generally obtain higher response rates than postal surveys.
- Evidence from a single study suggested that rates of item non-response may be higher for postal surveys.
- There is little consensus about the benefits of telephone and postal surveys on parameters of non-response bias, quality of response, anonymity or cost.

Self-completion questionnaires versus face-to-face interviews

- Face-to-face interviews tend to yield higher response rates.
- Evidence from a single study suggests that respondents may be more likely to give no answer at all or say "don't know" in an interview than in a self-completion questionnaire or card sort.

- There is a lack of unequivocal evidence to support the view that postal survey participants respond more truthfully to sensitive issues or make more critical or less socially acceptable responses than when face-to-face with an interviewer.

Telephone versus face-to-face interviews

- Telephone interviews may be quicker than face-to-face interviews.
- There is no consistent evidence of the relative superiority of face-to-face interviews or telephone interviews on the parameters of instrument response rate, eliciting sensitive information and item non-response rate.

Computer-assisted versus paper-based self-completion questionnaires

- Findings on the effects of computer-assisted versus paper-based questionnaires on response rates and response quality are equivocal.
- Evidence from a single study suggests that respondents to computerised questionnaires may use a wider range on rating scales.
- Quicker responses may be obtained by using computer-assisted questionnaires.
- There is no clear evidence that responses to sensitive questions differ between computer-assisted and paper-based modes.

Computer-assisted versus conventional telephone interviewing

- Interviewer variability may be lower with CATI.

General

- No one mode of administration consistently outperforms all others.

Recommendations for practice

Findings from high-grade primary studies were equivocal, suggesting that no single mode of administration is superior in all respects or in all settings. The choice of mode of administration should be therefore be made on a survey-by-survey basis, taking into account:

- study population
- survey topic
- sampling frame availability and quality
- sampling method
- volume of data to be collected
- complexity of data to be collected
- resources available.

Before embarking on a survey in a particular setting, with a particular population, or on a particular topic, the researcher should review carefully the literature to ascertain the appropriateness of the survey method in general and of different modes of survey administration in particular (including the likely impact of interviewer characteristics), in those particular circumstances.

Recommendations for future research

With the growing availability of and interest in information technology, priority should be given to comparative studies of traditional versus computer-assisted approaches, of different computer-assisted methods with each other, and of mixed-mode approaches, for example:

- computer-assisted interview approaches (CATI and CAPI) versus CASI
- traditional modes of data entry (data keying) from paper-based questionnaires versus electronic scanning of questionnaires (OMR and OCR)
- traditional keyboard entry for computer-assisted questionnaires versus more novel techniques such as touch-screen and light-pen data entry
- web-based delivery of questionnaires (particular issues here would be how to define and determine the underlying population and how to control for the same individual submitting multiple questionnaires)
- incorporation of traditional or computer-assisted self-completion segments into interviewer-administered surveys (e.g. to gather data on sensitive topics).

Particular attention should be paid to the relative merits of different modes of administration in surveys:

- of special populations (e.g. older people; ethnic communities; hearing-impaired people; motor-impaired people; health professionals)
- on sensitive topics (e.g. sexual behaviour; drug and alcohol use).

Future comparative studies of different modes of administration should use multiple outcome measures, including:

- the quantity of response (non-contact, ineligibility, refusal and instrument response rates; item non-response rates)
- the quality of response (non-response bias; validity, reliability and distribution of responses)
- resource implications (time to respond; cost per completed questionnaire).

TABLE 2 *Self-completion questionnaires versus telephone interviews*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin. (sample size)	Criteria for comparison	Main findings
Talley *et al.*, 1983[65]	RCT	3.5	Health: psychological needs	University students (USA)	Postal admin. (779) Telephone admin. (106)	Instrument response rates Rating of items	Instrument response rates: 35% postal; 79% telephone RR = 2.29 (95% CI, 2.00 to 2.63) telephone vs. postal Telephone respondents rated psychological needs as more important to them than postal respondents ($p < 0.0033$; insufficient data to calculate CIs) Items given most importance by telephone also given most importance by post; significant positive relationship found between 2 methods (Spearman's $\rho = 0.94$; $p < 0.0001$; insufficient data to calculate CIs)
McHorney *et al.*, 1994[66]	RCT	4	Health: national health survey using SF-36	Non-institutionalised adults (USA)	Postal admin. (2564) Telephone admin. (645) Telephone interviews were computer assisted; sample sizes refer to assigned mode of administration – for some respondents in each group, data had to be collected by the opposite method to that assigned	Instrument response rates Non-response bias Item non-response rates Reliability of responses Rating of items Cost	Overall instrument response rates: 79% postal; 69% telephone RR = 0.87 (95% CI, 0.82 to 0.92) telephone vs. postal Rate of response by assigned mode (i.e. mail assigned responding by mail etc.): 65% postal; 65% telephone RR = 1.00 (95% CI, 0.94 to 1.07) telephone vs. postal Non-response bias found in both modes, but was not differential apart from age: responders to postal survey significantly older than non-responders Mean no. missing responses (out of 104): 1.59 items postal; 0.49 items telephone Mean difference = 1.10 (95% CI, 0.70 to 1.50) Reliability of responses: internal consistency reliability of SF-36 scale scores range 0.63–0.93 (median 0.85) for postal and 0.77–0.93 (median 0.85) for telephone Health ratings less favourable for postal than telephone respondents Postal survey offered more anonymity for reporting sensitive and personal information Cost per completed questionnaire: $27.07 postal; $47.86 telephone

continued

TABLE 2 contd *Self-completion questionnaires versus telephone interviews*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin. (sample size)	Criteria for comparison	Main findings
Pederson *et al.*, 1994[67]	RCT	4	Health: attitudes to restriction on cigarette smoking in public places	Adults living in London, Ontario (Canada)	Postal admin. (58) [65] Telephone admin. (60) [87] [Nos contacted including substitutes]	Non-contact rates Refusal rates Instrument response rates Quality of response Cost	Non-contact rates: 11% postal; 43% telephone RR = 3.95 (95% CI, 1.88 to 8.29) telephone vs. postal Refusal rates: 34% postal; 11% telephone RR = 0.34 (95% CI, 0.17 to 0.67) telephone vs. postal Overall instrument response rates: 49% postal; 45% telephone RR = 0.91 (95% CI, 0.65 to 1.28) telephone vs. postal Instrument response rates for eligible respondents: 55% postal; 65% telephone RR = 1.18 (95% CI, 0.88 to 1.59) telephone vs. postal Cost per completed questionnaire: $34.06 postal; $28.74 telephone
Hinkle and King, 1978[64]	Non-random concurrent controlled study	3	Health: community mental health programme planning	Residents of Alabama (USA)	Postal admin. (1000) Telephone admin. (224) Interview admin. (641) In the interview group the survey was introduced by an interviewer, but the questionnaire was self-completed See also *Tables 3* and *4*	Non-contact rates Refusal rates Instrument response rates Achieved sample composition Rating of items Cost	Non-contact rates: 28% postal; 22% telephone; 22% interview RR = 0.80 (95% CI, 0.62 to 1.05) telephone vs. postal Refusal rates: 53% postal; 21% telephone; 8% interview RR = 0.33 (95% CI, 0.24 to 0.44) telephone vs. postal Instrument response rates: 19% postal; 57% telephone; 70% interview RR = 2.98 (95% CI, 2.51 to 3.54) telephone vs. postal All modes of admin. over-represented higher income, higher SES groups, but effect was most exaggerated for postal admin. More neutral and negative responses given by respondents to postal survey among higher SES users of mental health services Cost per completed questionnaire: $1.82 postal; $0.92 telephone; $1.34 interview

admin., administration; CI, confidence interval; RCT, randomised controlled trial; RR, relative risk; SES, socio-economic status; SF-36, Medical Outcomes Study Short Form-36 questionnaire

TABLE 3 *Self-completion questionnaires versus face-to-face interviews*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin. (sample size)	Criteria for comparison	Main findings
Newton *et al.*, 1982[68]	RCT	3	Health: feelings about social life	Adults aged ≥18 years (USA)	(1522 in total) All 1522 respondents completed approximately 67 items (randomly allocated) by each mode of admin.: self-completion; card sort; interview	Item non-response rate Rating of items	Item non-response rates significantly different across modes ($p < 0.05$; insufficient data to calculate CIs) for 84/202 (42%) of items Respondents more likely to say "don't know" in interview than in self-completion questionnaire or card sort Mean difference significantly different in ratings across modes ($p < 0.05$; insufficient data to calculate CIs) for 46/202 (23%) items Interview had highest mean score on 21/46 items Respondents more willing to endorse a negative item or reject a positive item in interview compared with self-completion questionnaire or card sort Items high in social desirability no more subject to form effects than items low in social desirability Form effects not correlated with respondents' social background
Oei and Zwart, 1986[70]	RCT (cross-over)	4	Health: life events	Psychiatric hospital inpatients (The Netherlands)	(58 in total) All respondents completed both a self-completion questionnaire and an interview; the order of admin. was randomised in a cross-over design	Rating of items (reporting of significant life events)	Mean no. life events: 4.20 self-completion; 2.74 interview Mean difference = 1.5 (95% CI, 0.5 to 2.5) Death reported more often in interview mode; working conditions, education/training, illness (self) and marital problems reported more often in self-completion mode
Cartwright, 1988[71]	RCT (alternate allocation after random selection)	3.5	Health: women's experiences of maternity services	New mothers (UK)	Postal admin. (400) Interview admin. (400)	Instrument response rates Non-response bias Item non-response rates Nature of responses	Instrument response rates: 75% postal; 92% interview RR = 1.23 (95% CI, 1.15 to 1.31) interview vs. postal Non-response bias: Asian mothers under-represented among those responding to postal survey (insufficient detail to calculate significance levels or CIs) Item non-response rate means: 1.9% postal; 0.6% interview No difference in nature of responses; replies to painful or delicate subjects similar in both groups; some support for view that criticisms would be reported at interview rather than in postal questionnaire

continued

TABLE 3 contd *Self-completion questionnaires versus face-to-face interviews*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin. (sample size)	Criteria for comparison	Main findings
Boekeloo *et al.*, 1994[73]	RCT	3	Health: reports of HIV risks	Hospital outpatients (USA)	Audio self-completion (153) Written self-completion (152) Both groups then received face-to-face interview; results relate to comparison	Item non-response rates Rating of items (reporting of risky health behaviour)	Item non-response rate means: 3.0% audio self-admin.; 6.4% written self-admin.; 0.0% interview (insufficient data to calculate significance levels or CIs) For 6/16 items, risky behaviour more frequently reported in audio self-completion questionnaire than in interview; for 4/16 items, risky behaviour more frequently reported in written self-completion questionnaire than in interview
Hinkle and King, 1978[64]	Non-random concurrent controlled study	3	Health: community mental health programme planning	Residents of Alabama (USA)	Postal admin. (1000) Telephone admin. (224) Interview admin. (641) In the interview group the survey was introduced by an interviewer, but the questionnaire was self-completed See also *Tables 2* and *4*	Non-contact rates Refusal rates Instrument response rates Achieved sample composition Rating of items Cost	Non-contact rates: 28% postal; 22% telephone; 22% interview RR = 0.79 (95% CI, 0.66 to 0.94) interview vs. postal Refusal rates: 53% postal; 21% telephone; 8% interview RR = 0.15 (95% CI, 0.11 to 0.20) telephone interview vs. postal Instrument response rates: 19% postal; 57% telephone; 70% interview RR = 3.69 (95% CI, 3.21 to 4.23) interview vs. postal All modes of admin. over-represented higher income, higher SES groups, but effect was most exaggerated for postal admin. More neutral and negative responses given by respondents to postal survey among higher SES users of mental health services Cost per completed questionnaire: $1.82 postal; $0.92 telephone; $1.34 interview
Liefeld, 1988[72]	RCT	2	Non-health: views of visitors to 2 shopping malls	Shoppers (Canada)	Self-completion (261) Computer-assisted (239) Interview (288) Figures refer to achieved sample sizes; there were refusals to participate from 2 computer-assisted admin.; failure to complete occurred in 5 self-completion, 10 computer-assisted admin. and 2 personal interviews See also *Table 5*	Response patterns	For fact-type questions (e.g. those requiring a "yes/no" response), few significant response differences found between 3 methods For multiresponse questions testing knowledge, computer-assisted questionnaire respondents picked more incorrect answers than personal interview and self-completion respondents Greatest differences in responses were between computer-assisted and self-completion methods; differences much smaller between computer-assisted and personal interview (Analysed by ANOVA; insufficient data to conduct pair-wise comparison of methods, or to calculate CIs)

continued

TABLE 3 contd *Self-completion questionnaires versus face-to-face interviews*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin. (sample size)	Criteria for comparison	Main findings
Nederhof, 1984[69]	RCT	3	Non-health: hypothetical division of money between participant and second party	Inhabitants of a medium-sized town (The Netherlands)	Postal admin. (108) Interview admin. (108) Postal group was further subdivided: 54 prenotified by post and 54 by telephone	Rating of items (provision of socially desirable responses)	Mean score on altruistic and socially desirable items: 3.62 postal–postal; 3.83 telephone–postal; 3.36 interview ($p < 0.002$; insufficient detail to calculate CIs) (lower score is more socially desirable)

ANOVA, analysis of variance

TABLE 4 *Telephone interviews versus face-to face interviews*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin. (sample size)	Criteria for comparison	Main findings
Jordon *et al.*, 1980[74]	RCT	2.5	Health: health beliefs	Los Angeles County residents (USA)	Telephone admin. (1114) [613] Face-to-face admin. (2020) [1883] [No. eligible subjects, upon which all other calculations based]	Ineligibility rates Instrument response rates Item non-response rates Rating of items Quantity of responses to open-ended and checklist questions	Ineligibility rates: 36% telephone; 7% face-to-face RR = 0.19 (95% CI, 0.16 to 0.23) face-to-face vs. telephone Instrument response rates: 49% telephone; 64% face-to-face RR = 1.30 (95% CI, 1.19 to 1.42) face-to-face vs. telephone Missing data on family income: 21% telephone; 12% face-to-face RR = 0.57 (95% CI, 0.47 to 0.69) face-to-face vs. telephone For Likert scale questions, telephone admin. led to: More acquiescence: mean difference = –1.4 (95% CI, –1.9 to –0.8) Greater evasiveness: mean difference = –0.1 (95% CI, –0.2 to 0.0) More contradictions: mean difference = –1.6 (95% CI, –1.7 to –1.5) No significant effect of mode of admin. on no. responses to open-ended questions (insufficient data to calculate CIs) More responses to multiple response checklist questions for telephone admin. (insufficient data to calculate CIs)

continued

TABLE 4 contd *Telephone interviews versus face-to face interviews*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin. (sample size)	Criteria for comparison	Main findings
Aneshensel *et al.*, 1982[76]	RCT	3	Health: reports of physical morbidity	Residents of Los Angeles County aged ≥18 years (USA)	Telephone admin. (377) Face-to-face admin. (296)	Instrument response rates Reported health	Instrument response rates: 82% telephone; 80% face-to-face RR = 0.98 (95% CI, 0.91 to 1.06) face-to-face vs. telephone Reported bed-disability days in 2 weeks prior to interview: 6% telephone; 11% face-to-face RR = 2.06 (95% CI, 1.15 to 3.68) face-to-face vs. telephone Reported other restricted-activity days in 2 weeks prior to interview:11% telephone; 18% face-to-face RR = 1.59 (95% CI, 1.05 to 2.40) face-to-face vs. telephone Reported restricted activity (composite variable) in 2 weeks prior to interview: 13% telephone; 24% face-to-face RR = 1.77 (95% CI, 1.23 to 2.55) face-to-face vs. telephone No other significant differences in reported health
Hinkle and King, 1978[64]	Non-random concurrent controlled study	3	Health: community mental health programme planning	Residents of Alabama (USA)	Postal admin. (1000) Telephone admin. (224) Face-to face admin. (641) In the interview group the survey was introduced by an interviewer, but the questionnaire was self-completed See also *Tables 2* and *3*	Non-contact rates Refusal rates Instrument response rates Achieved sample composition Rating of items Cost	Non-contact rates: 28% postal; 22% telephone; 22% interview RR = 0.99 (95% CI, 0.74 to 1.31) face-to-face vs. telephone Refusal rates: 53% postal; 21% telephone; 8% interview RR = 0.38 (95% CI, 0.26 to 0.55) face-to-face vs telephone Instrument response rates: 19% postal; 57% telephone; 70% interview RR = 1.24 (95% CI, 1.09 to 1.40) face-to-face vs. telephone All modes of admin. over-represented higher income, higher SES groups, but effect was most exaggerated for postal admin. More neutral and negative responses given by respondents to postal survey among higher SES users of mental health services Cost/completed questionnaire: $1.82 postal; $0.92 telephone; $1.34 interview
Fenig *et al.*, 1993[77]	Non-random concurrent controlled study	3.5	Health: psychiatric survey of those who had and had not experienced the Holocaust	Index group: women who had experienced the Holocaust Comparison group (matched for sociodemographic variables): women who resided in prestate Israel during the Holocaust period (Israel)	153 approached, 145 agreed to participate (76 index group; 69 comparison group) All participants approached for telephone interview, 115 (randomly selected) subsequently approached for face-to-face interview	Instrument response rates Reliability of responses Detection of significant demoralisation Rating of items	Instrument response rates: 95% telephone; 83% face-to-face RR = 0.88 (95% CI, 0.81 to 0.96) face-to-face vs. telephone Reliability (internal consistency): Cronbach's α = 0.94 telephone; 0.90 face-to-face Mean score on "Rate of demoralisation" scale: 1.84 telephone; 1.45 face-to-face index group ($p < 0.001$; insufficient data to calculate CI) 1.15 telephone; 1.06 face-to-face control group (p = ns; insufficient data to calculate CIs) After Bonferroni correction for multiple comparisons, 27 items were significantly higher ($p \leq 0.001$; insufficient data to calculate CIs) for telephone admin.

continued

TABLE 4 contd *Telephone interviews versus face-to face interviews*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin. (sample size)	Criteria for comparison	Main findings
Quinn *et al.*, 1980[75]	RCT	2.5	Non-health: information on mundane and sensitive topics	Residents of Ann Arbor, MI, whose names were in the telephone directory (USA)	Telephone admin.: (425) [343] Face-to-face admin.: (398) [347] [No. eligible respondents upon which all other calculations based]	Instrument response rates Refusal rates Achieved sample composition Item non-response rates Quality of data (no. "problems" in life domains reported) Interviewers' estimates of respondents' reactions Interviewers' personal satisfaction with interview Length of interview	Ineligibility rates: 19% telephone; 13% face-to-face RR = 0.66 (95% CI, 0.48 to 0.92) face-to-face vs. telephone Refusal rates: 24% telephone; 18% face-to-face RR = 0.74 (95% CI, 0.55 to 0.99) face-to-face vs. telephone Instrument response rates: 72% telephone; 68% face-to-face RR = 0.94 (95% CI, 0.86 to 1.04) face-to-face vs. telephone Achieved sample composition (% female): 70% telephone; 55% face-to-face ($p < 0.05$; insufficient data to calculate CI) Mean no. "not ascertained" responses: 3.36 telephone; 3.29 face-to-face (p = ns; insufficient data to calculate CI) Mean no. responses to open-ended questions assigned to residual coding categories: 1.04 telephone; 1.04 face-to-face (p = ns; insufficient data to calculate CI) No. "problems" reported: 3.01 telephone; 3.02 face-to-face (p = ns; insufficient data to calculate CI) Interviewers' estimation of respondents' reactions to interview (higher scores denote more favourable attitudes): Interest: 2.65 telephone; 2.77 face-to-face ($p < 0.05$; insufficient data to calculate CI) Lack of suspicion: 2.73 telephone; 2.85 face-to-face ($p < 0.001$; insufficient data to calculate CI) No significant effect of mode of admin. on: rapport, honesty or comprehension Interviewers' personal satisfaction with interview: 2.70 telephone; 2.77 face-to-face face ($p < 0.05$; insufficient data to calculate CI) Length of interview (mean no. min): 20.9 telephone; 24.0 face-to-face ($p < 0.001$; insufficient data to calculate CI)

ns, not significant

TABLE 5 *Computer-assisted versus paper-based self-completion questionnaires*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin. (sample size)	Criteria for comparison	Main findings
Allen, 1987[78]	RCT	2.5	Non-health: experiences of university	Students (USA)	Paper-based admin. (124) Computer-assisted admin. (125) Paper-based questionnaires were designed for direct scanning (OMR) to computer	Instrument response rates Achieved sample composition Response distribution Response validity Respondent evaluation	Response rates: 49% paper-based; 29% computer-assisted RR = 0.60 (95% CI, 0.43 to 0.83) computer-assisted vs. paper-based No significant differences in race, gender, grade point average, degree topic, status or residence type in achieved samples (insufficient data to calculate CIs) Response distribution: computer-assisted admin. produced a higher SD and used a significantly wider range of rating scales than paper-based admin. for 6/12 items No consistent differences in validity of reported responses (in comparison with university records) Respondent evaluation (% who would recommend survey to a friend): 85% paper-based; 97% computer-assisted RR = 1.14 (95% CI, 1.01 to 1.29) computer-assisted vs. paper-based No significant difference with respect to evaluation of convenience: paper-based group believed survey to be "about right length"; computer-assisted group "too short" ($p < 0.001$; insufficient data to calculate CI)
Helgeson and Ursic, 1989[80]	RCT	3	Non-health: views on brands of products in fast-food restaurant	Undergraduate business students (USA)	Paper-based admin. (60) Computer-assisted admin. (60) (Also manipulated length of questionnaire) See also *Table 9*	Rating (means and variances of items) Decision-making processes	Overall mean rating: 2.72 paper-based; 2.88 computer-assisted (p = ns; insufficient data to calculate CI) Overall variance of rating: 1.22 paper-based; 1.20 computer-assisted (p = ns; insufficient data to calculate CI) Decision-making processes: no consistent differences in processing codes between the two modes of admin.

continued

TABLE 5 contd *Computer-assisted versus paper-based self-completion questionnaires*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin. (sample size)	Criteria for comparison	Main findings
Higgins *et al.*, 1987[79]	RCT	3.5	Non-health: views of work-at-home methods	IBM PC users (Canada)	Paper-based admin. (103 approx.) Computer-assisted admin. (103 approx.) Both modes (103) A total of 308 individuals were randomised to 3 groups, receiving respectively: paper-based questionnaire only; disk-based questionnaire only; both paper-based and disk-based questionnaires. For the purposes of analysis, the 36 respondents in the third group who returned a paper-based questionnaire were assigned to the "paper-based" group and the 36 who returned a disk-based questionnaire were assigned to the "disk-based" group	Instrument response rates Achieved sample composition Item non-response rates Quantity of response Response bias Estimated vs. actual completion time Response speed	Response rates (raw): 63% paper-based; 66% disk-based RR = 1.05 (95% CI, 0.88 to 1.24) disk-based vs. paper-based Response rates (adjusted for inability to participate): 63% paper-based; 78% disk-based RR = 1.21 (95% CI, 1.03 to 1.42) disk-based vs. paper-based No significant differences (insufficient data to calculate CIs) in achieved samples with respect to 9 demographic variables, self-reported computer skills and knowledge of other people receiving the same questionnaire Item non-response rates (% not responding to income questions): 13% paper-based; 19% disk-based (p = ns; insufficient data to calculate CI) Mean words in response to open-ended questions: 31.02 paper-based; 39.24 disk-based (p = 0.03; insufficient data to calculate CI) Mean no. points raised in response to open-ended questions: 5.84 paper-based; 6.89 disk-based (p = 0.01; insufficient data to calculate CI) Response bias (mean scores "sensitive" questions – higher score is "more honest"): 11.86 paper-based; 11.00 disk-based (p = ns; insufficient data to calculate CI) Completion time (mean no. min) for disk-based questionnaire: estimated 23.12; actual 30.18 (p < 0.001; insufficient data to calculate CI) Response speed (mean days to respond): 6.68 DISKQ; 8.85 paper (p < 0.001; insufficient data to calculate CI)

continued

TABLE 5 contd *Computer-assisted versus paper-based self-completion questionnaires*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin. (sample size)	Criteria for comparison	Main findings
Liefeld, 1988[72]	RCT	2	Non-health: views of visitors to 2 shopping malls	Shoppers (Canada)	Self-completion (261) Computer-assisted admin. (239) Interview admin. (288) Figures refer to achieved sample sizes; there were refusals to participate from 2 of those assigned to computer-assisted admin.; failure to complete occurred for 5 assigned to self-completion admin., 10 computer-assisted admin. and 2 personal interviews See also *Table 3*	Response patterns	For fact-type questions (e.g. those requiring a "yes/no" response), few significant response differences were found between 3 methods For multiresponse questions testing knowledge, computer-assisted questionnaire respondents picked more incorrect answers than personal interview and self-completion respondents Greatest differences in responses were between computer-assisted and self-completion methods; differences were much smaller between computer-assisted and personal interview (Analysis was by ANOVA; insufficient data to conduct pair-wise comparison of methods, or to calculate CIs)

SD, standard deviation

Chapter 4

Question wording and sequencing

Introduction

In this chapter, research evidence concerning a major aspect of what Miller[82] described as the "rich folklore of survey research" – namely, question wording and sequencing, and the wording and ordering of question response formats – are examined and synthesised. In doing so the authors recognise that recommendations cannot be made in the abstract; rather, they must take into account theories and empirical findings regarding respondent behaviour.

Moser and Kalton[5] considered question design to be "largely a matter of art rather than science", in which "common sense and past experience are the surveyor's main tools". They nonetheless proposed a number of guiding principles that need to be taken into account in posing questions, including:

- Respondents will give some kind of answer to most questions, even if they are ill-informed, and will offer opinions on matters to which they have given little thought.
- Response accuracy will be limited by memory errors.
- Faced with sensitive or threatening questions, respondents may mislead, understate or exaggerate.
- Respondents' attitudes may be latent, many-sided, inconsistent and of variable strength.

Drawing on these principles, they cautioned against the use of questions that:

- are insufficiently specific
- are hypothetical
- employ technical words and jargon
- are leading or presuming
- are vague or ambiguous.

With regard to the order of questions, Moser and Kalton[5] proposed that the researcher should:

- start with the easier questions and work through to the more difficult
- order the questions in a logical sequence and leave personal demographic questions to the end (because these may be sensitive)
- design questionnaires so that one question or group of questions sets the context for later ones or, conversely, so that respondents' answers to later questions are not influenced by those preceding them.

The art of asking questions was further addressed by Sudman and Bradburn,[7] who divided them into two classes: those that are, in principle, verifiable (behavioural and factual questions); and those that are not verifiable even in principle (psychological state or attitudinal questions). They also distinguished between questions that are threatening to the respondent and those that are not, and suggested strategies for handling each type. In two earlier books, the same authors[8,26] had concluded that question structure and length do not affect response effects for non-threatening questions. However, in "threatening" questions asking about the frequency of socially undesirable behaviour, closed questions (i.e. forced or multiple choice) were shown to increase the likelihood of under-reporting in comparison with open-ended questions. Similarly, shorter questions were more likely to result in under-reporting than longer questions. However, these effects were not observed in questions that asked simply whether the socially undesirable activity had been carried out or not. Furthermore, contrary to expectations, a more familiar form of words (e.g. in which respondents were allowed to supply their own term for "intoxication") was not shown to be superior to standardised wording.

For non-threatening questions about behaviour, Sudman and Bradburn[7] proposed that:

- Questions should be as specific as possible.
- All reasonable response alternatives should be included.
- The time-frame should be related to how salient or memorable the topic is.
- The use of aided-recall procedures and memory cues should be considered.
- Permission to consult documentary sources may be given.

Their recommended techniques for obtaining accurate responses to threatening questions included:

- using long introductions
- using open-ended questions

- using familiar or colloquial words
- using an appropriate time-frame
- deliberately loading the question towards the reporting of socially undesirable behaviour, by:
  - indicating that the behaviour is very common (e.g. "Most people occasionally go to bed without cleaning their teeth. How many times in the last week have you done this?")
  - assuming the behaviour and asking merely about frequency or other details (e.g. "How many cigarettes do you smoke each day?" with the option of responding "None")
  - citing authority to justify behaviour (e.g. "Many doctors now say that drinking red wine reduces the risk of heart disease. Have you drunk any red wine in the past month?")
- embedding the threatening topic within a list of more and less threatening subjects to reduce its perceived importance
- using techniques such as card sorting and randomised response.[5]

To optimise responses to knowledge questions, Sudman and Bradburn[7] suggested:

- using filter questions to screen out respondents who lack sufficient information
- including a "don't know" response category to reduce the perceived threat
- using open-ended questions even though numerical answers are required (to counter any tendency to choose the mid-point)
- using pictures and other non-verbal procedures as well as standard questions
- asking several questions on the same topic to reduce the likelihood of successful guessing especially where "yes/no" responses are required.

Finally, with regard to attitude questions, they proposed that:

- Double-barrelled questions should be avoided.
- Questions should be standardised by explicit specification of the alternatives.
- A middle response category should be included unless there are persuasive reasons not to do so.
- General questions seem to be more susceptible to ordering effects, so they should be asked before specific ones.
- If measuring changes in attitudes over time, exactly the same questions should be asked at each time point.

With regard to response formats, particularly when measuring attitudes, Sudman and Bradburn[7] suggested that:

- Open-ended questions should be used only sparingly (because they are more resource intensive and are more subject to inter-interviewer variability).
- Response categories should start with the least socially desirable option.
- Rating scales should be limited to not more than five points when written descriptors are attached.
- For more than five response categories, numerical scales should be used.
- Analogues such as ladders, clocks or thermometers should be considered for numerical scales that have many points.
- Respondents should be asked to respond to every item in a list rather than indicating only those that apply (i.e. to respond "yes/no" or "applies/does not apply" to each item rather than simply complying with an instruction such as "circle as many as apply").

Regarding question ordering, Sudman and Bradburn[7] proposed that:

- Easy, salient and non-threatening questions should come first in a questionnaire.
- Demographic questions, because they can be seen as threatening, should come last, unless required to screen for eligibility.
- In interview surveys, funnelling procedures should be used in order to minimise question order effects, starting with the general and moving to the specific.
- Questions on one topic should be completed before embarking on a new topic.
- Transitional phrases and instructions should be used when switching topics.
- Filter questions should be ordered in such a way as to cover all contingencies and encourage complete responses.

Many of the issues addressed in these two standard texts and outlined above have been the subject of further scrutiny; indeed, several books have been written on the subject.[8,26,83–86] Relevant findings from primary studies identified in the course of this review are discussed below under three broad headings: question wording, question sequencing and response format.

Question wording

Sudman and Bradburn[7] commented that a badly worded questionnaire, "like an awkward conversation, can turn an initially pleasant situation into a boring or frustrating experience". Most respondents want to give the best information

they can, so it is incumbent on the researcher to facilitate this process by developing questions that are clearly formulated and precise. They should try to address four factors related to response error:

- memory
- motivation
- communication
- knowledge.

Identified studies

Eleven studies were identified (*Table 6*; see p. 63) that met the quality criteria and addressed the issue of question wording.[87–97] All were randomised controlled trials; however, the survey topic was health related in only two[87,95] of the 11, which may limit the generalisability of the findings to health surveys. Six of the surveys were interviewer administered (telephone interviews in five cases), one used a postal questionnaire, and three used self-completion questionnaires with captive audiences; the mode of administration was not specified in one case. The quality of the studies was generally high, although none reported explicitly that sample size calculations were based on a power calculation. For a number of them, the data presented were insufficient to allow us to compute 95% CIs for all the findings. In other articles, the number of multiple comparisons precluded the presentation of RRs (or mean differences) and associated 95% CIs in the tables of results.

The specific issues covered by the identified studies were:

- open-ended versus closed questions[93]
- one- versus two-sided attitudinal questions[90]
- direct versus indirect questions[88]
- elliptical versus non-elliptical question wording[96]
- time-framing in question wording[94,95]
- negative versus positive, or neutral versus non-neutral questions[87,89,90]
- wording of filter question[91,92]
- use of prestige names in questions.[97]

Use of open-ended versus closed questions

The limitations of form and wording of closed questions was examined by Schuman and colleagues,[93] who investigated attitudes of the US population to a range of social issues, including the threat of nuclear war, in a series of five telephone surveys. Random halves of each survey sample were asked either to respond to an open-ended question ("What do you think is the most important problem facing this country today?"), or to a closed question ("Which of the following do you think is the most important problem facing this country today: the high cost of living, unemployment, the threat of nuclear war, or government budget cuts, or, if you prefer, you may name a different problem as the most important."). Although the open-ended question produced a larger range of different responses than did the closed version, Schuman and colleagues[93] noted that most of these led to the creation of additional categories that were either very small or else "vaguely miscellaneous" in nature. Furthermore, for four of the five surveys, the rankings of four "common issues" were identical on the open-ended and closed question forms, suggesting that, as long as response categories to a closed question include the main issues identified by an open-ended question (in a pilot study), the use of either will ordinarily lead to similar conclusions.

These observations highlight the importance of adequate development work and piloting of a questionnaire, using open-ended questions to identify the most important issues for inclusion as response categories in closed questions. Furthermore, the observation that the closed form used by Schuman and colleagues[93] elicited responses other than those explicitly offered, suggests the desirability of including a category of "other, please specify" in closed questions.

One- versus two-sided attitudinal questions

Bishop and co-workers[90] investigated the effects of presenting one or two sides of an issue in survey questions. Respondents to an omnibus telephone survey conducted in a major metropolitan area of the USA were randomly assigned to either a one-sided presentation of an issue in agree/disagree format (e.g. "Now some people are afraid that the Government in Washington is getting too powerful for the good of this country and the individual person. Do you agree or disagree with the idea that the Government is getting too strong for the good of the country and the individual person?") or to a two-sided presentation in forced-choice form (e.g. "Now some people are afraid that the Government in Washington is getting too powerful for the good of this country and the individual person. Others feel that the government in Washington has not gotten too strong for the good of the country and the individual person ... What is your feeling: do you think the Government is getting too powerful or do you think the government has not gotten too strong?"). On two out of five issues, presenting a second substantive choice stimulated significantly higher percentages of respondents to give an opinion. Comparing the percentage of respondents agreeing with a one-sided statement in agree/disagree format with the

percentage selecting the same statement over a second alternative in a two-sided substantive choice format showed that offering the second alternative decreased endorsement of the initial single-sided statement. These findings suggest that question wording and format may affect both whether or not informants will offer an opinion on a topic and what opinion they will give.

Direct versus indirect questions

Salvendy[88] investigated the effect on survey response rates of framing questions in direct form or indirectly in the form of a short story that enabled the respondent to imagine himself/herself in a specific situation and to give the appropriate response (analogous to the "vignette" approach sometimes used in health-related research). In both versions, "yes/no" responses were required to each of 16 questions. No overall differences in response rates were found. However, the results indicated that people with higher education were as willing to respond to indirect as to direct questions, but this was not the case for those with lower educational levels.

Elliptical versus non-elliptical question wording

In the field of linguistics, sentences that are verb- or noun-less (i.e. shortened forms of other sentences, for example "What?..." or "How come?..."), are referred to as elliptical sentences.[98] In such sentences, the listener's "linguistic competence" enables him or her to interpret the meaning through the application of certain "transformational" rules.[99] One study[96] that met the review's quality criteria examined the effect on survey responses of including what the authors considered to be both elliptical and non-elliptical structure questions. Different versions of a 15-item questionnaire were constructed such that in one version the so-termed elliptical structure questions (e.g. "Advertising leads to wasteful buying in our society.") appeared before the non-elliptical equivalent (e.g. "I think advertising leads to wasteful buying in our society."); in the other, the reverse was the case. The two versions were administered to randomly assigned groups of adult volunteers. There was no difference in survey responses as a result of the ordering of the questions, but the analysis showed that more than a fifth of respondents answered differently to the two question forms; the elliptical structure questions produced more polarised responses than the non-elliptical. Subtle changes in question wording may therefore cause shifts in response patterns, and combining both question forms may introduce bias into the results.

Time-frames in question wording

Behavioural frequency questions occur commonly in health surveys and the accuracy with which participants respond to them is therefore a subject of concern. For threatening or sensitive questions, motivational considerations are seen as the key in explaining response error;[100] for non-threatening questions, memory errors represent the greatest threat.[26] Sudman and Bradburn[26] proposed two sources of memory error in non-threatening questions: episode omission, whereby the respondent fails to recall an event falling within the specified time-frame; and episode telescoping, whereby the respondent misplaces an event in time (either by placing it more recently in time or more distant in time than is really the case). A range of methods for reducing these types of errors has been proposed. For example, Sudman and Bradburn[26] recommended aided or cued recall, to be achieved by the inclusion of "for example ..." prompts, and the use of cue cards. Sudman and Ferber[101] and Wind and Lerner[102] suggested that diaries could be of use, while Neter and Waksberg[103] advocated the use of bounded recall; this latter technique involves repeated interviews with the same panel of respondents, in which the earlier interviews are used to set boundaries on the recall period for later interviews.

The explanations proposed by Sudman and Bradburn[26] about possible sources of memory error in non-threatening questions rest on the underlying assumption that respondents recall and count relevant behavioural episodes in formulating answers to survey questions involving frequency of behaviour (i.e. employ "episodic enumeration"). However, it is possible that other cognitive processes may also be involved. For example, Blair and Burton[94] suggested that respondents may use a heuristic, based on rate of occurrence, to derive estimates of frequency. (This suggestion is supported by the experiences of one of the authors (EMcC) in developing a questionnaire to measure the frequency of symptoms in patients with asthma. When asked how they had arrived at their estimates of symptom frequency over the preceding 3 months, several respondents reported thinking about the rate of occurrence in the past month and extrapolating from that rate.)

Blair and Burton[94] argued that a clearer understanding of the underlying cognitive processes would allow those carrying out surveys to phrase and administer questions in such a way as to reduce response error. Accordingly, they investigated, by making the process of episodic enumeration more complex, whether increasing the specified time-frame in behavioural frequency questions reduces the accuracy of reporting of events, and whether the use of two alternative

question cues influences the process of enumeration. In a 3 × 2 factorial study design, they examined the effects of specified time-frames for recall ("2 weeks", "2 months" or "6 months") and question format (either "how many times" or "how often" a particular behaviour had occurred). Respondents to a telephone survey were asked about the frequency of six behaviours (all non-health related). At the end of each interview, the interviewers questioned respondents about the cognitive processes they had employed in formulating their answers. There was a significant association between reported use of episodic enumeration processes and question time-frame, with this type of enumeration being more common for shorter time-frames. However, there was no difference between use of the cues "how often" and "how many times" in reported application of episodic enumeration processes. The answers to the question regarding cognitive processes ("How did you come up with that answer?", asked in respect of a question on frequency of dining out) highlighted that episodic enumeration represented only one of 12 distinct cognitive processes employed by respondents in arriving at their estimates of behavioural frequency; in fact, episodic enumeration accounted for only 28% of the answers. The most commonly reported cognitive process was rate-based estimates (estimation of a rate of behavioural activity without any recall of specific behavioural episodes), but episodic enumeration was more commonly used by those reporting a low frequency of the target activity. Blair and Burton[94] concluded that existing models of the way in which respondents formulate their answers to survey questions are incomplete and so may be inadequate.

A related study by Larsen and colleagues[95] looked at the use of the quantifiers "frequently" and "often" in question wording. They conducted two experiments to investigate reported experience of headache, the first involving a sample of college students, the second a general population sample. In the first study, two versions of a ten-item health questionnaire were distributed to random halves of the sample, one containing a question that asked if they experienced headaches "frequently", the other whether they experienced headaches "occasionally". In the second study, eight versions of a health questionnaire containing questions about frequency of headaches and tooth-brushing, in which the quantifiers varied, were randomly distributed to men and women attending a public event. In this latter study, Larsen and colleagues[95] compared the reported frequency of headaches in response to two separate questions ("Do you get headaches frequently (occasionally)? If so, how often?") and to a single question ("How frequently (often) do you get headaches?"). Among the student sample, there was no significant difference in the mean number of headaches reported between the two forms of wording; however the quantifier "frequently" led to a significantly higher percentage saying they did not get headaches. Similarly, in the general public sample, respondents were more likely to say "no" to the first of the two questions when the quantifier "frequently" was used, but the overall mean number of headaches reported did not differ significantly across the four versions of the headache questions. Nor was there any difference in the reported frequency of brushing teeth in response to questions regarding "How frequently …" or "How often …". Larsen and colleagues[95] concluded that the quantifier "frequently" may lead to underestimates of the overall prevalence of headaches in a question requiring a "yes/no" response. They cited the example of a woman who had headaches approximately once a week, but answered "no" to the question using "frequently" because she considered her headaches to be relatively uncommon.

Positively versus negatively worded, or neutral versus non-neutral questions

Received psychometric wisdom on the subject of respondent acquiescence is that attitudinal measures should contain an even balance of positively and negatively worded items, to avoid what has been termed "response set" bias, whereby respondents simply endorse the same (numbered) response category for each item.

Schriesheim and Hill[89] used three questionnaire formats (one with all items positively worded, one with all items negatively worded, and one with a mixture of positively and negatively worded items), to examine acquiescence response bias ("yea-saying") among business administration students. Participants were each asked to read a one-page script describing behaviours displayed by a fictitious supervisor before being randomly assigned to receive one or other of the three questionnaires. These authors then computed "accuracy scores", which measured the discrepancies between the students' descriptions based on their questionnaire responses and the researchers' own judgement based on the description provided in the script. The use of all positively worded items resulted in "more accurate" descriptions by this criterion than did the use of either mixed or all negatively worded ones. The decrements in accuracy appeared to be a function of the negatively worded items themselves, rather

than of their exerting a strong contextual effect on the positive items. Schriesheim and Hill[89] concluded that the inclusion of negative items in attitudinal questionnaires may impair rather than increase the validity of survey results.

As part of the omnibus survey described above, Bishop and colleagues[90] investigated the effect of positively versus negatively worded single-sided questions on respondent acquiescence (i.e. the tendency to agree with the statement). Thirty-six per cent of respondents agreed with the positively worded statement (i.e. government should "see to it") compared with 26% who disagreed with the negatively worded one (i.e. government should "stay out of it"), representing what Bishop and colleagues[90] termed an "acquiescence effect" of 10% ($p < 0.05$). This finding suggests that estimates of the strength of agreement with an issue may be affected by whether the related statement is worded positively or negatively. Of course, it is not possible to say which approach produced the most valid data. The likelihood of acquiescing was significantly related to educational level, rising from 2% for college-educated respondents to 25% for those with less than high school education. Bishop and colleagues[90] argued that resolution of these issues requires the development of an information-processing model in which the responses of informants to questionnaires are recognised to be a function of the information available to them within a specific context; they suggested that the development of such a model is preferable to methodological refinements with regard to question wording and format.

One health-related study on the impact of neutral rather than non-neutral instructions was identified.[87] A random half of women volunteers who were asked to complete the Moos Menstrual Distress Questionnaire received the original version, which specified that the symptoms listed were menstrual; the rest received a "masked" version, where no cause for the symptoms was assigned in the accompanying instructions. There were no significant differences between the two groups in the mean number of symptoms reported either overall or at any stage of the menstrual cycle (premenstrual, menstrual, intermenstrual), indicating that the condition-specific (i.e. non-neutral) instruction did not encourage participants to report symptoms stereotypically rather than as they were actually experienced.

Wording of filter question

As already noted, Sudman and Bradburn[7] have suggested that, to optimise responses to knowledge questions, filter questions should be used to screen out respondents who lack sufficient information.

In a series of multipurpose, random-digit-dialled telephone surveys, Bishop and colleagues[92] showed that the pressure on informants to give answers to fictitious questions can be reduced by the introduction of an explicit filter question that allows them to indicate that they have given little thought to the topic under investigation. In an earlier study the same research team[91] also examined the impact of differently worded filter questions. The effect of a filter was shown to depend on both the content of the item and the wording of the filter. Willingness to give a "don't know" response was found to vary with the strength of the filter question (i.e. the degree of encour-agement to respond in this way). The more "remote" the question topic was from the respondent's experience (as reflected by the greater frequency of "don't know" responses elicited on a questionnaire without filters), the greater the effect of adding filters. Finally, Bishop and colleagues[91] also concluded that the presence and nature of filter questions can also significantly affect the distribution of substantive responses (i.e. those expressing a definite opinion) in such a way as to alter the conclusions drawn from them.

Use of prestige names in questions

It is possible that the use of "prestige" names (those implying certain values or points of view) in a question represent a subtle form of loading. One study was identified that examined the effects of using prestige names in question wording.[97] In two surveys on political topics, the inclusion of a prestige name (that of a controversial politician) markedly reduced the number of respondents having no opinion and added a partisan component (i.e. a different political mix among those offering opinions, potentially a source of bias) to the questions; this latter effect was greatest when respondents knew least about the subject matter being investigated. The study results confirm conventional survey wisdom that prestige names should, wherever possible, be excluded from questions because they represent additional stimuli and so additional sources of variability between respondents, which, in turn, compromises the interpretation of responses.

Question sequencing

The effect of question ordering on the responses given to questions has been widely investigated and it has been shown that placing an item in different positions in a questionnaire may alter the way in

which respondents answer it. It has been suggested that placing sensitive, unpleasant or embarrassing questions early in an interview schedule may increase the likelihood of informants answering in what they consider to be a socially acceptable way or refusing to answer at all. Question ordering may be subject to a "consistency effect", whereby responses to a given question are brought into line with responses to earlier questions.[104] A general recommendation is that questions asking for a general evaluation should precede more specific questions because the latter may create a "saliency effect", which influences answers to more general questions.[104] If question order influences the nature of responses, then altering it in repeat surveys may hamper the interpretation of any observed differences in response patterns over time because these may be an artefact of question context rather than an indication of true change. It has been argued that question order effects will be minimal in postal compared with interview surveys[105] because of reduced serial-order constraints and the fact that respondents are not forced by time constraints to give "top-of-the-head" responses;[106] furthermore, in self-completion questionnaires, participants have the opportunity to read all the questions before responding.

Identified studies

Fourteen studies (*Table 7*; see p. 69) were identified on the topic of question sequencing. Of these, three were health-related randomised controlled trials,[107–109] nine were randomised trials on non-health topics[110–118] and two were health-related historically controlled studies.[119,120] Nine involved interviewer-adminis-tration (in eight cases by telephone), four were postal surveys and one a self-completion survey conducted in a workplace. As with those on question wording, the quality of the studies was generally high, although none reported explicitly that sample size calculations were based on a power calculation. For a number of the studies the data presented were insufficient to allow calculation of 95% CIs for all the findings. In other articles the number of multiple comparisons precluded the presentation of RRs (or mean differences) and associated 95% CIs in the table of results.

The specific issues addressed were:

- general context effects[108,110,112,114,115,117,118,120]
- ordering of general versus specific questions[107,111,117,119]
- contiguous versus non-contiguous questions[113,116]
- ordering of disease-specific versus generic health status instruments.[109]

General context effects

Two studies (both non-health related) examined the effect of question ordering on overall response rates. Jones and Lang[110] examined context effects as one of a series of possible factors (the others being survey sponsorship, wording of the covering letter, and method of notification) affecting overall response rates to a postal survey on a non-health topic. In the first version of the questionnaire a set of 42 semantic-differential attribute scales appeared before a set of 27 anchored similarity-judgement scales; in the second version the positions of these two sets were reversed. In both versions these scales were preceded by demographic questions. The semantic-differential attributes items were printed on a single page; although this gave a dense appearance, it was considered that each judgement could be made quite easily and therefore this section of the questionnaire could be completed quickly. The anchored similarity-judgement scales were also printed on a single page, but each judgement represented a more complicated evaluation task. The first version resulted in significantly higher response rates than the second but was also associated with increased sample composition bias (with respect to house purchase prices among respondents compared with those for the entire sampling frame). Jones and Lang[110] commented that the effects of sample composition bias on survey precision may be substantial.

Roberson and Sundstrom[118] also reported on the effects of topic order on overall response. In their self-completion employee attitude survey, the ordering of six topics and of a series of demographic questions was manipulated in a 6×2 factorial design. A "prioritised" order, based on the rankings of employee representatives, produced a higher response rate than each of five random orders; so did location of the demographic questions at the end rather than the beginning of the questionnaire. Topic order, but not demographic item location, also significantly affected attitudinal responses. The prioritised topic order reflected employees' expressed concerns, so Roberson and Sundstrom[118] concluded that one of the most important aspects of questionnaire design relates to the early items; once respondents have been encouraged by the apparent relevance of these items to embark on the task of filling in the questionnaire, they are more likely to complete it.

Two health and four non-health studies addressed the issue of general context effects on question responses. Context effects were examined in a telephone survey by Colasanto and colleagues[108] in

relation to questions about AIDS infection. They looked at the outcome of placing a question about whether AIDS could be transmitted through blood donation before or after a question on blood transfusion as a means of transmission. Respondents were more likely to indicate a belief that it is possible to contract AIDS through blood donation when this question preceded the one on blood transfusion. It appeared that a preceding question on blood transfusion helped to clarify the meaning of the potentially ambiguous donation question by means of a "contrast" effect[84] and so reduced the number of erroneous responses.

Using a historically controlled study design, Serdula and colleagues[120] examined context effects in relation to another health question, weight loss. Data were derived from a series of telephone surveys conducted annually between 1985 and 1992, involving just under 250,000 respondents. From 1985 to 1988 the respondents were first asked what their body weight was and then whether they were currently trying to lose weight; in the subsequent surveys the order of the two questions was reversed. Forty-eight per cent of women and 29% of men in the first series, and 41% of women and 26% of men in the second series, reported that they were trying to lose weight; prevalence differences were 7.1% (95% CI, 6.3 to 7.9) and 2.3% respectively (95% CI, 1.4 to 3.2). Serdula and colleagues[120] concluded that survey respondents, particularly women, are more likely to report dieting when questions about weight control practices follow questions on current weight. Such question-context effects may therefore bias prevalence estimates and invalidate comparisons across surveys where the same questions are asked but not in identical order. An alternative explanation of a true historical effect – that the prevalence of dieting had indeed increased over time – was also considered by Serdula and colleagues;[120] however, comparative data from the National Health Interview Surveys[120] showed no change in the proportion of people attempting weight loss between 1985 and 1990, suggesting that there was no historical effect.

A study by Sigelman[112] considered the impact of placing a question about presidential popularity at the beginning or near the end of a telephone interview schedule. In the version with the question at the end, it was preceded by a series of "negatively charged" questions about social issues. It was hypothesised that this latter ordering would lead to less favourable ratings of presidential popularity. However, Sigelman[112] found no significant difference in the direction of evaluations between the two versions, but there was a statistically significant difference in the proportions of respondents willing to offer any evaluation of the president. This effect was more pronounced for less well-educated respondents, of whom 20% expressed no opinion when the question was asked first, compared with only 8% when it was asked last ($p < 0.01$). Thus the effect of question order on willingness to express an opinion may pose a significant potential threat to the comparability of survey results across populations or over time.

Spector and Michaels[114] examined the hypothesis that question order effects may also threaten the validity of research findings that are based on self-report when satisfaction and perceptual questions are included in the same questionnaire. To investigate this possibility they compared results from two versions of a postal organisational questionnaire. In one, job satisfaction questions preceded questions on job perceptions; in the other, the order of these sections was reversed. Out of 300 possible comparisons, only 13 were significantly different, leading these authors to conclude that, at least within the context of organisational research, such order effects are not an important problem.

Tourangeau and colleagues[115] explored the theory that respondents' answers to attitude questions involve the retrieval of their beliefs that are relevant to those questions, and that this retrieval process is in turn affected by stimuli in the form of prior questions. In survey settings, prior items in a questionnaire serve to "prime" respondents by temporarily altering the belief retrieval process. If these items trigger their retrieval of beliefs relevant to the target item, a "carry-over" effect will occur. To investigate this possibility, respondents in a telephone survey were interviewed twice about a number of social issues, including abortion, welfare spending and defence. In the second interview, items that preceded the target item on each issue were varied systematically. When the context items were "more favourable" in relation to the target items, respondents gave responses to the target questions that were more consistent with their responses to the related context items, supporting the hypothesis about belief accessibility. Tourangeau and colleagues[116] also reported that context effects are most marked for respondents who indicate that their beliefs about a target issue are both mixed and important to them.

The authors of one study attempted to examine whether the question order effects manifested in interview surveys also exist in postal surveys.

Ayidiya and McClendon[117] argued that the observation of question order effects in interview surveys may be attributable either to respondents' internal need to be consistent in their answers (i.e. to appear consistent to themselves) or to their desire to present a consistent image to the interviewer. If the latter is the case, then such effects should, in theory, be eliminated in self-completion surveys. To test this hypothesis, they used a postal questionnaire to collect data from a systematic sample of 532 US households. These researchers detected a question order effect for one issue (about a Communist reporter working in the USA and an American reporter working in Russia), which, although smaller than in previous interview surveys (e.g. Schuman and colleagues[113]), was nonetheless in the same direction. However, for another issue (abortion), on which question order effects were detected in previous interview studies,[107,119] no such effect was found, a finding echoing that of Bishop and colleagues.[121] Ayidiya and McClendon[117] concluded that their original hypothesis could have been too simplistic and that it may therefore be necessary to specify more carefully the types of question order effects that may occur in self-completion and postal surveys.

Ordering of general versus specific questions

Three of the identified studies[107,117,119] examined ordering effects with respect to a general and a specific question about abortion.

In the study by Schuman and colleagues,[107] initial data were collected as part of a national telephone survey in the USA. The general abortion item ("Do you think it should be possible for a pregnant woman to obtain a legal abortion if she is married and does not want any more children?") received significantly more support when asked before the specific item ("Do you think it should be possible for a woman to obtain a legal abortion if there is a chance of serious defect in the baby?") than when asked after the specific one. However, the percentage of respondents who replied positively to the specific item was unaffected by its ordering. These findings were replicated in a second survey by the same researchers, which also indicated that the order effects were not influenced by respondent characteristics (gender, religious affiliation or educational level). The order effect was more marked for those who professed themselves to be undecided on the issue of abortion than for the others, suggesting a greater resistance to such effects among respondents who feel strongly about a particular issue. Schuman and colleagues[107] commented that context effects may be especially likely when the investigator attempts to summarise complex issues with a single general item that fails to make allowance for any qualifications or ambivalence on the part of respondents.

This view was endorsed by Tenvergert and colleagues,[119] who compared retrospectively, using a historically controlled design, levels of endorsement of one general and two specific questions about abortion in three general public interview surveys carried out in the USA in 1984, 1987 and 1988. They reported that approval of the general item (about whether a married woman who did not want any more children should have access to legal abortion) was significantly higher in the 1987 survey; on this occasion, the general question was asked before the specific item on the availability of abortion in the likelihood of a defective baby but after that on the availability of abortion if the health of the mother was at risk. However, in contrast with the earlier work by Schuman and colleagues,[107] Tenvergert and co-workers[119] also found an order effect for the specific item on a defective baby; this item was more likely to be positively endorsed when placed after the general question. However, no such order effect was found for the other specific question on the health of the mother.

Ayidiya and McClendon[117] also tested the abortion questions in a postal survey. In contrast to the other researchers, they did not identify any significant ordering effect, a finding echoing that of Bishop and colleagues in a self-administered survey.[121]

A non-health survey in which the order effects of general versus specific questions were addressed was carried out by McFarland,[111] who examined variations in responses associated with placing general attitudinal questions to a range of social issues before or after a series of specific ones. A random sample of 516 respondents were interviewed by telephone and, in four non-overlapping sections, were asked about their attitudes to energy, the economy, politics and religion. Each section of the questionnaire contained one general question and a series of specific ones. Respondents were found to be significantly more likely to express an interest in politics and religion when the general question followed the specific, although their general evaluations on the other two topics were unaffected by question order. There was no evidence that the strength of order effects varied with respondent characteristics (gender and education level). In the light of these findings, McFarland[111] concluded that, although some questions may be more susceptible than others, and order effects are not necessarily ubiquitous,

the customary recommendation that general questions should precede specific ones appears to be justified.

Contiguous versus non-contiguous questions ("blocking" and "buffering")

Schuman and colleagues[113] examined whether interposing neutral items between questions known or thought likely to influence one another could reduce question order effects. In a telephone survey, the focus of which was US–USSR relations, the interview schedule included a pair of items shown in previous research to be clearly susceptible to context effects (these items regarded the freedom of Communist reporters to report from the USA and US reporters to report from Russia). In the first and second versions of the schedule the items were placed contiguously but with their order reversed; in the third they were placed non-contiguously, being separated by 17 other (mainly demographic) questions. The context effect was confirmed and shown to be only slightly (and non-significantly) reduced by the separation of the two items of interest by the neutral ones. Schuman and colleagues[113] concluded that, although the non-contiguous positioning of items may counter context effects where these effects are weak, it is unlikely to do so where they are pronounced.

Another important finding from the study by Tourangeau and colleagues[115] described under general context effects above, and which supports earlier research findings, was that context effects were reduced by the introduction of irrelevant buffer items. Tourangeau and colleagues explored this phenomenon further in a second study,[116] in which over 1000 respondents to a telephone survey were asked about six target (social and political) issues. There were ten different versions of the questionnaire. All started with 20 attitude questions on issues not closely related to the target issues. The next section contained the six questions on the target issues and four questions on each of six related context issues. Two separate sets of context issues were used (e.g. US involvement in Lebanon or international terrorism in Iran for the target issue of involvement in the Persian Gulf). In four versions of the questionnaire, the items in context set 1 were presented before the target items; in a further four, the items in context set 2 preceded the context items; the remaining two versions used context set 1, with these questions positioned after the target items. For four of the eight versions in which the context items preceded the target items, the context items appeared immediately before the related target items ("blocked" versions), while, in the others, the context items were scattered ("scattered" or "buffered" versions). The ordering of the target items was also varied across the different versions of the questionnaire. A significant "context group" effect ($p < 0.001$) was found; those whose questionnaires included context set 1 gave responses to the target items that were "more consistent" with their answers to the context items. Although there was an overall trend towards greater context effects when context items were "blocked" immediately before the related target item, this effect was weak and not statistically significant. However, based on a meta-analysis of their own results with those from four other studies addressing the same issue (including the one by Schuman and colleagues[113] described above), Tourangeau and co-workers[116] concluded that buffering items may significantly reduce, but are unlikely completely to eliminate, context effects.

Ordering of disease-specific versus generic health status instruments

There is considerable emphasis in the health outcomes literature on the value of combining disease-specific and generic instruments in health status questionnaires,[122] so the question of their ordering within questionnaires is highly relevant. One health-related randomised study[109] that addressed this issue was identified. Men with symptomatic benign prostatic hyperplasia were prospectively enrolled into a clinical trial of an educational intervention and assigned to one of two versions of a baseline questionnaire. In one version, a 38-item disease-specific question module appeared first, followed by a 36-item generic health status module, the SF-36; in the other, the position of the modules was reversed. Scores were compared for the three disease-specific subscales and the eight subscales of the SF-36. There were no statistically significant differences between the two versions in the distribution of scores across any of the disease-specific or generic subscales. Neither were there any differences between the two versions in the magni-tude of correlation coeffi-cients between the disease-specific and generic subscales. These results suggested no effect of instrument ordering. However, Barry and colleagues[109] highlighted a number of limitations in their study, including that its focus was on only one disease condition with "a relatively bounded impact on overall health status"; they concluded that larger studies in other disease states are required to determine whether their results are generalisable across other outcomes research.

Response format

Sudman and Bradburn[7] made the point that, to some extent, the distinction between question

wording and response formats is an artificial one because the form of the question often dictates the most appropriate response format.

Identified studies

Fourteen studies[82,92,117,123–133] that dealt with the issue of response format were identified (*Table 8*; see p. 75). All but one were randomised controlled trials. The remaining study was a non-random concurrent controlled study in which the two groups were equivalent with regard to age, gender and reported difficulty with completing the questionnaire. All except two[82,123] of the studies were non-health related and 11 used postal or self-completion questionnaires. As with the studies of question wording and sequencing, quality scores were affected by a lack of power calculations in determining sample size and by reporting findings in insufficient detail to compute RRs, mean differences and CIs. Once again, space constraints precluded the presentation of parameter estimates and CIs where there were extensive multiple comparisons.

The specific aspects of response category construction and presentation addressed were:

- inclusion of "no opinion" or "don't know" categories, and "middle" responses[92,117,123,127,132,133]
- ordering of response categories[82,117,126,128]
- labelling of response categories[117,125,129,130]
- remote versus adjacent scale placement[124]
- including a space for free comment at the end of a list of attitudinal items employing a closed response format.[131]

Inclusion of "no opinion", "don't know" and "middle" responses

The question of whether to include or exclude "no opinion", "don't know" or "middle/neutral" responses in survey question response options is one that has been considerably debated and each viewpoint has been subject to both criticism and support. The inclusion of these response options may mean, particularly in the context of self-administered questionnaires, that respondents opt for an easy way out rather than taking time to think about their attitudes or record factual information; however, their exclusion may encourage respondents to make wild guesses or omit questions altogether. Six articles meeting the quality criteria addressed this issue.

In the context of telephone surveys, Bishop[132] reported the effects of including a "middle alternative" response category (generally meaning a continuation of the *status quo*). He examined specifically the effects of: offering versus omitting from the list an explicit middle alternative of response categories read to the respondent (in the "omit" version, a middle alternative response was accepted if volunteered by the respondent); including the middle alternative in the question while omitting it from the response categories offered; and the position in which the "middle" alternative was presented (as the second or last of three response categories). The questions related to three US public policy issues: social security benefits, defence spending and nuclear power plants. As hypothesised by Bishop,[132] respondents were much more likely to choose the middle alternative when it was offered explicitly than when it was not. On two of the three issues investigated, participants were also more likely to select the middle alternative when it was mentioned in the preface to the question even though it was not offered as an explicit response category (e.g. "Some people believe we should spend less money on defence. Others feel that defence spending should be increased. Still others feel that defence spending should be continued at the present level. How about you – do you think that defence spending should be increased or decreased?"). The order in which the "middle" response was offered (second or last position in the list) also affected response distributions; although the pattern of effects was not consistent or invariable, there was a tendency towards increased endorsement of the "middle" alternative when it was presented as the last response category (possibly a "recency" effect[52]). Bishop[132] suggested that respondents are more likely to opt for the "middle" response category when it is offered if the polar opposite alternatives are equally attractive or unattractive, making the choice between them problematic. Finally, in 5/12 comparisons, the distribution of polar responses (i.e. those coming down on one side or the other) were significantly different between the versions offering and omitting a middle alternative; the difference approached statistical significance for two other comparisons. This finding suggests that conclusions derived from a survey may be affected by the inclusion or otherwise of a middle alternative.

This issue was further investigated by Wandzilak and colleagues,[133] who hypothesised that the responses of people who selected the middle response category ("undecided") in one form of a self-completion questionnaire would be equally distributed across the adjacent polar categories ("agree" and "disagree") when they completed a second form in which the mid-point was omitted. This hypothesis was not supported; responses were skewed on all eight items considered and the

favouring of one side over the other was statistically significant for three of the eight items. Wandzilak and colleagues[133] concluded that the middle response category does not necessarily represent a position of neutrality and its exclusion either at the data collection or analysis stage may produce inaccurate results.

Ayidiya and McClendon[117] tested the effect of offering an explicit middle alternative in a postal survey. When this alternative ("middle of the road" with respect to political issues) was offered explicitly, the percentage of respondents who described their political attitudes in this way was significantly higher than when only the polar alternatives ("liberal" or "conservative") were offered explicitly and respondents had to volunteer "middle of the road" choices. However, when those choosing the middle alternative were omitted, the distribution of polar responses did not vary significantly between the two forms. "Primacy" effects[52] were observed for one out of three comparisons and there was no evidence of "recency" effects.[52]

The studies by Bishop,[132] Wandzilak and colleagues,[133] and Ayidiya and McClendon[117] described above involved attitude questions. Poe and colleagues[123] evaluated the effects of including or excluding an explicit "don't know" box in a postal questionnaire comprising only factual questions, which was sent to the close relatives of a sample of recently deceased persons in the USA. In the version with the "don't know" boxes, respondents were instructed to check this box, or to write "?" in the answer space if they did not know the answer to a particular question. In the version without the "don't know" boxes, respondents were simply instructed to write "?" in the answer space. For the purposes of analysis, any indication by respondents that they did not know the answer was coded as "don't know". Poe and colleagues[123] reported that overall rates of return of a completed questionnaire were unaffected by the presence or otherwise of "don't know" boxes, as was the average percentage of items left blank. However, the average percentage of "not known" ("don't know" or "?") responses was significantly higher for the version with "don't know" boxes. For a quarter of the items in the questionnaire, the percentage of substantive (i.e. usable) replies was significantly higher for the version without "don't know" boxes. For most items the distribution of substantive responses did not differ significantly between the two versions. For half of those items where a significant difference in response distributions was observed, the differences in percentage endorsement of specific substantive response categories were less than 10%. Based on these findings, Poe and colleagues[123] concluded that self-administered questionnaires without "don't know" boxes are to be preferred for a number of reasons: the absence of these boxes did not affect overall response rate; without them, the questionnaire looked less cluttered and skip instructions were easier to follow; an appreciably higher rate of useable responses was engendered; less imputation of missing responses was required; and, importantly, there were few differences in response distributions between the two versions.

Two studies examined the question of whether a "don't know" category influenced responses to a fictitious issue. Although health surveys are unlikely to include questions on unauthentic topics, they may address issues of which respondents have little knowledge or experience, but which they may be tempted to answer in an effort to be helpful.

Hawkins and Coney,[127] in a mail survey of lawyers and the lay public in the USA, reported no impact of a "don't know" response option on overall response rate or on response to questions about which respondents were informed. However, the inclusion of a "don't know" option appeared to reduce the rate of uninformed responses (i.e. the expression of an opinion on a fictitious issue). These authors therefore recommended that such an option should be provided unless there are explicit reasons not to do so.

Bishop and colleagues[92] also investigated the effect of a "don't know" response option on respondents' willingness to offer an opinion on a fictitious issue, this time within the context of an interview. Three different versions of the questionnaire were used. In the first, a filter question was initially used to screen out those respondents who had not thought much about the issue; in the second, interviewers did not probe those who indicated that they did not know anything about the topic or gave a response other than the two options offered (e.g. "agree" or "disagree"); in the third, if respondents volunteered a "don't know" response, interviewers pressed for a definitive answer. On all three fictitious issues about which they were questioned (none of which was health-related), respondents in the third group were more likely to express agreement or disagreement. Bishop and colleagues[92] concurred with Hawkins and Coney[127] that offering respondents an explicit opportunity to have "no opinion" helps to avoid the creation of a spurious form of representativeness.

Finally, Ayidiya and McClendon[117] experimented with a "no opinion" filter for four questions in a

postal survey. For all four questions the percentage responding "don't know" to the target questions was significantly higher for the filtered form.

Ordering and presentation of response categories

Response order effects in closed survey questions are said to occur when the order in which the response alternatives are listed influences respondents' choices.[84] Two types of response order effects that can arise in surveys and have been the subject of examination are "primacy" effects (the tendency to select the first appropriate response category) and "recency" effects (the tendency to select the last relevant response option). According to Schuman and Presser[84] and to Krosnick,[52] primacy effects occur more frequently in self-completion (including postal) surveys and in interview surveys where cue cards are used, while recency effects are more likely in interview surveys (including telephone interviews) in which the respondent simply listens to a list of choices read out by the interviewer.

Israel and Taylor[128] examined general response order effects in questions requiring single and multiple responses and reported evidence of their occurrence in the latter, although not in the former. For one of three multiple response questions, the percentage endorsing response category A was significantly higher when this category was presented first rather than last; however, the percentage endorsing the other five categories offered for this question was not affected by the order of presentation. For a further multiple choice question involving attribution (and therefore believed to be prone to social desirability bias – the tendency to select the response perceived to be socially acceptable), the percentage endorsing the "socially acceptable" category was significantly higher when this was offered first rather than third; once again, the percentage endorsing each of the other categories for this item was not affected by the order in which they were presented. Israel and Taylor[128] made the point that, since the order of response alternatives may affect not only the distribution of responses to individual items but also, as a consequence, associations between these and other items, ignoring the possibility of response category ordering effects runs the risk of estimating incorrectly the effects of an intervention.

Ayidiya and McClendon[117] investigated the extent of primacy and recency effects in postal surveys in relation to three questions on housing, morality and divorce. They reported a marginally significant primacy effect ($p = 0.08$) on the question about divorce, with the *status quo* category (ability to obtain a divorce should "stay as it is now") selected more often when presented in the middle position than in the last position. Strictly speaking, this is not a primacy effect as defined above, but it is indicative of a greater tendency to select a response category the earlier it is presented. There were no statistically significant recency effects across the three questions, supporting previous evidence that these are uncommon in self-administered surveys.[84,121]

Edvardsson[126] studied the effect of reversing response categories across six different groups of questions in a self-completion questionnaire distributed to a group of psychology students. In version A, negative responses were listed before positive responses (e.g. scale ordered: "not at all interested" – "somewhat interested" – "fairly interested" – "very interested"); in version B, positive responses appeared before negative responses. The number of scale points varied from set to set, ranging from two ("yes" – "no") to five ("totally disagree" – "hesitantly disagree" – "undecided" – "hesitantly agree" – "totally agree"). No statistically significant differences were found in response distributions across a total of 53 items or for any of the six sets. However, Edvardsson[126] cautioned the reader about the limits of generalisability of the results, given the population studied (university students), the form and content of the items, and the types of response scales.

One health-related study[82] compared approaches to asking attitudinal questions in telephone interviews, in which visual prompts for the response categories are absent. In one arm of the study, 7-point attitude scales were presented to respondents in a single step; more specifically, they were asked to indicate which number on a 7-point scale best described their satisfaction with various aspects of their health. In the other, a two-step approach was used, whereby respondents were first asked for a general statement of attitude (i.e. "satisfied" or "dissatisfied") and then for a more detailed specification (i.e. "completely", "mostly" or "somewhat" satisfied, or dissatisfied). Those who responded "in the middle" to the first query were probed to see if they were leaning to one side or were truly ambivalent. Although the response formats were not strictly comparable because the one-step process asked for a numerical response and the two-step process for a verbal one, Miller[82] argued that to make them so would be operationally problematic because this would involve interviewers in the one-step process in reading out and repeating all possible verbal response options. (It is because of this lack of

comparability, in addition to the lack of a power calculation and the fact that results are not reported in sufficient detail to allow the calculation of CIs, that this study was given a quality score of 2.) Despite a tendency for the two-step approach to produce higher mean scores, the difference reached statistical significance for only one out of the five questions. However, these non-significant differences in means masked variations in the pattern of responses. Higher proportions of those receiving the two-step approach declared themselves to be "completely satisfied"; more of those receiving the one-step approach chose the neutral, middle response; the one-step approach also led to higher endorsement of the number 5 category ("somewhat satisfied"), while the two-step approach led to relatively greater endorsement of the number 3 category ("somewhat dissatisfied"); the two-step approach consistently yielded higher levels of missing data. Although none of these differences in individual categories was "large by visual inspection" (according to Miller), for each of the five questions the differences in response distributions were statistically significant. Based on the finding of broadly similar mean scores, Miller[82] concluded that the two approaches were largely interchangeable, suggesting that this refuted the argument that attitudinal questions should be broken down into components for telephone administration. However, he went on to express a preference for one-step administration because: it produced fewer selections of the most positive category; it produced less missing data; the items showed higher correlations with each other; and inter-viewers found this version easier to administer. The present authors' own belief is that the two approaches have not been demonstrated by Miller[82] to yield equivalent results. However, in the absence of a "gold standard" measure of satisfaction, it is not possible to tell which one yields the more valid data.

Labelling of response categories

Anchor point effects were examined by Frisbie and Brandenburg[125] and by Lam and Klockars.[130]

Frisbie and Brandeburg[125] queried whether respondents focus on the verbal descriptors (e.g. "excellent" to "poor") or on the numerical codes on labels attached to each category (e.g. the numerals 1–5). They compared scales (both 4- and 5-point) in which only the end-points had a verbal descriptor attached with those in which all scale points were labelled with a descriptor. When only the scale end-points were labelled, mean scores for six of the eight questions were higher (more favourable); the non-significant differences were for one 5-point scale ("very good" – "good" – "satisfactory" – "needs improvement" – "very poor") and one 4-point scale ("strongly disagree" – "disagree" – "agree" – "strongly agree"). In this study, the authors[125] also considered the issue of labelling response categories numerically (i.e. 1–5) or alphabetically (i.e. A–E) and found no significant differences between the two. Thus, although item equivalence can be assumed for alphabetical and numerical labelling, it appeared, from this study, that the same is not true for end-point-defined versus all-point-defined scales. Bipolar adjective scales (i.e. those with only the end-points verbally described) may yield higher ratings than fully defined scales; the use of the latter may counteract leniency errors (the tendency to bias ratings in an upward direction). However, Frisbie and Brandenburg[125] noted that the observed differences in their study may not have been of practical significance (they ranged in magnitude from one-third to one-ninth of a standard deviation).

Lam and Klockars[130] speculated that the findings of Frisbie and Brandenburg[125] may be specific to the set of verbal descriptors used in that study. They hypothesised that the relationship between responses to scales in which only the end-points are labelled and to fully labelled scales is a function of the descriptors attached to the intermediate points. They compared responses to questions on four types of 5-point rating scales. On one, only the end-points were verbally described ("poor" – "excellent"); on the other three, all five points were described, but with different response configurations between the two end-points. Lam and Klockars[130] referred to these as "equally spaced" ("poor" – "needs improvement" – "satisfactory" – "quite good" – "excellent"); "positively packed" ("poor" – "fair" – "good" – "very good" – "excellent"); and "negatively packed" ("poor" – "moderately poor" – "fair" – "good" – "excellent"). Mean responses were approximately equal for the scale in which only the end-points were labelled and the "equally spaced" scale; both were located between the means for the "negatively packed" and "positively packed" scales. This finding of no significant difference between the "end-points only labelled" and "equally spaced" scales, which contradicted the results from Frisbie and Brandenburg,[125] suggested to Lam and Klockars[130] that the earlier result was an artefact of the descriptors used. They also concluded that respondents pay close attention to the content of response alternatives when answering questionnaire items and that they interpret undefined intermediate alternatives by mentally dividing the

scale into equal intervals. They suggested that it may therefore be sufficient to describe the end-points only. They also suggested that finer discrimination within a limited part of the scale can be achieved, if so desired, by "packing" the scale with response options from one part of the underlying continuum.

Neither of these studies was in the context of health and their conflicting conclusions regarding scale-point labelling suggest that further work is required to investigate the importance of anchor points in responses to items in health surveys.

The labelling of response categories has also been investigated in relation to sensitive questions in mail questionnaires. Swan and Epley[129] tested the hypothesis that using wide rather than narrow response categories in questions relating to income would increase both overall questionnaire response rates and item completion rates. They found that category width had no influence on questionnaire response rates, with equal percentages of those receiving wide- and narrow-income-band questionnaires returning them. Nor did it influence significantly the item completion rates. Swan and Epley[129] also examined the effect of allowing respondents to endorse two adjacent response categories rather than a single one as a means of providing less precise information about income. They found a slight but non-significant difference in return rates and a significant difference in question completion rates. Contrary to expectations, item completion rates were higher for the version requiring the endorsement of one category only. On the basis of these findings, Swan and Epley[129] recommended the use of narrow response categories in questions on income if the "information objectives" of the survey require high precision.

Ayidiya and McClendon[117] noted that, when asked whether they agree or disagree with a given statement (e.g. "Individuals are more to blame than social conditions for crime and lawlessness in this country."), some respondents have a tendency to agree with the statement more often than if it is presented with the same statement in a "forced-choice" format (e.g. "Which in your opinion is more to blame for crime and lawlessness in this country – individuals or social conditions?"). This acquiescence effect is believed to be more common among less well-educated respondents, possibly due to less sophisticated cognitive processing or to a desire to defer to interviewers who are perceived to be of higher status. Ayidiya and McClendon[117] tested for the presence of an acquiescence effect in a postal survey. For three separate questions, the percentage agreeing with the statement was higher in the "agree–disagree" format. The differences were statistically significant ($p < 0.05$) for two of the three items; however, there was no reliable evidence that acquiescence effects were greater among the less well-educated respondents.

Other issues relating to response format

The authors of one identified study considered not the ordering of response categories in relation to one another but their positioning adjacent to or remote from the question stem.[124] In the remote scale format, the rating scale (i.e. the response categories) was shown at the top of each page or new section of the questionnaire, requiring that the respondents referred back to this point as they answered each question. In the adjacent scale format, the rating scale appeared alongside each question. Stem and colleagues[124] found that the remote scale format tended to elicit more neutral responses. They speculated that this finding may have been the result of respondents becoming less confident as they moved further away from the rating scale. They suggested that this tendency towards neutrality could be minimised by repeating the scale after every three or four questions. Nonetheless, they concluded that remote and adjacent scale formats may yield different results.

Finally, one article was identified[131] that examined the effect on response rates of including a space for free comments at the end of a series of attitudinal questions using Likert-type responses. Response rates were doubled for questionnaires with a space for comments compared with those without. More comments were elicited when the space for their insertion was unstructured than when areas for consideration were specified to the respondent.

Conclusions

Some caution is required in extrapolating the findings of previous research to health surveys. Few of the identified studies were health related; the generalisability of findings to this field may be limited. For many of the topics investigated it was possible to identify only one or two relevant studies that met the quality criteria. Moreover, theories of both cognition and response formulation, as well as empirical evidence, suggest that the effects of question and response category wording and ordering may vary with the mode of administration, again indicating a need for caution in interpreting

and applying findings from interview surveys to self-completion questionnaires or vice versa.

Question wording

- Question wording and format can influence both whether or not an opinion is given, and what opinion is given.
- Open-ended questions produce more non-common category responses than closed questions, but most additional categories are small and miscellaneous. The use of either question form will ordinarily lead to similar conclusions. However, expert opinion suggests that open-ended questions still remain important in development stages and pilot studies; using open-ended questions at these stages allows researchers to generate appropriate response categories for closed questions.
- Survey participants employ a wide range of cognitive processes in formulating responses to behavioural frequency questions, including episodic enumeration (i.e. recalling and counting specific instances on which the behaviour occurred) and rate processing (i.e. aggregating from the "normal" rate at which the behaviour takes place in a "convenient" unit of time, such as a week). Task conditions, such as the time-framing of a question, will influence the processes employed.
- The wording of filter questions asking whether the respondent has knowledge of or has thought about an issue can affect significantly the percentages of "don't know" responses elicited to a subsequent substantive question, particularly for topics that are less familiar to the informant. Conversely, the content of the question can have an important independent effect on "don't know" responses, regardless of the filter wording. The use of filter questions can alter the conclusions drawn.
- Giving a second substantive choice on attitudinal questions increases the likelihood of respondents expressing an opinion.
- Acquiescence effects tend to be negatively related to the educational level of the respondents.
- Response bias may be introduced by the use of mixed grammar chains (e.g. elliptical versus non-elliptical structures). Elliptical structure questions (those in which the verb is omitted) produce both more agreement and more disagreement, while non-elliptical ones produce more neutral responses.
- Although the inclusion of negatively phrased items may theoretically control or offset acquiescence tendencies, their actual effect may be to reduce response validity.
- The interpretation of questions that include prestige names is complicated by the fact that participants respond not only on the basis of the content of issues but also on the basis of the names. Prestige names represent both additional stimuli and additional sources of variance to be explained.

Question sequencing

- Question order effects may influence overall response rates and increase sample composition bias in a range of ways; the direction and strength of these effects can vary with the topic, context and study population.
- The design of a questionnaire, particularly the apparent relevance of opening items, may influence people's motivation to complete the instrument; the more salient and relevant these items are, the greater the likelihood of response.
- The received wisdom that questions should be grouped by topic and ordered so that related topics are adjacent to one another ignores the issue of context effects; yet research suggests that such effects are common. Researchers need to balance the risk of context effects with the desirability of coherence and continuity.
- Topic ordering within a questionnaire may differentially affect response rates among different attitudinal groups.
- Question order effects may not be ubiquitous, but evidence suggests that general questions should precede specific ones.
- Context effects may bias estimates of the prevalence of attitudes and behaviour. If otherwise identical questions are posed in a different order or a different context across questionnaires (either at the same point in time or in longitudinal studies), apparent differences in response patterns may reflect context effects rather than true differences between respondents.
- Context effects are especially likely when researchers attempt to summarise complex issues in a single general item.
- Context effects may be larger when respondents' beliefs about a target issue are both mixed and important to them.
- Prior items in a questionnaire may exert a "carry-over" effect by priming respondents about their beliefs/attitudes towards a particular topic.
- "Buffering" of items may reduce context effects but is unlikely to eliminate them completely.
- Question order effects tend to be consistent across gender and educational levels and so are as much of a concern in surveys on restricted populations as in those on general populations.
- Context effects may be lessened but not entirely eliminated in self-completion questionnaires.
- Evidence suggests that scores on disease-specific

and generic health status measures in health outcomes questionnaires are unaffected by their position relative to one another. This question was found to have been investigated in only one disease area with a relatively bounded impact on overall health status, so further studies are required to determine whether the results are generalisable.

Response format

- The inclusion of middle position, no opinion and "don't know" response options seems generally preferable for attitudinal questions, although they may be less important for factual ones.
- Providing informants with an opportunity to have no opinion may avoid spurious representativeness.
- The "middle response" category does not necessarily represent a position of neutrality and its exclusion may produce invalid results.
- The wording of response categories is as critical as question wording because ambiguity in their meaning contributes to response order effects.
- The order of response alternatives may affect both the distribution of responses to individual items and associations between these and other items.
- Recency effects (the tendency to choose the last response option) appear uncommon in self-completion questionnaires.
- Findings on the relative merits of single-step and two-step approaches to presenting response categories for attitudinal questions in telephone surveys are equivocal.
- Evidence about the labelling of response categories is inconsistent, but fully defining scales may act as a check on leniency errors.
- A remote scale format in which the response categories are at a distance from the question appears to be associated with a tendency towards neutrality of response.
- The inclusion of a space for free comment may increase response rates.

Recommendations for practice

Recommendations with an evidence base from one or more high-grade primary comparative studies

Question wording

- Efforts to increase response accuracy should take into account the range of cognitive processes involved in response formulation and the potential impact of task variables such as: the likely salience and temporal regularity of events; the method of survey administration; and question design issues such as the time-frame. (Recommendation based on evidence from primary research studies and on theories of response formulation.)
- Open-ended questions should be used sparingly, particularly in self-completion questionnaires. However, careful piloting and pretesting by using open-ended questions should be carried out to ensure that the response categories presented in closed questions adequately represent the likely range of responses. (Recommendation based on evidence from primary research studies and on expert opinion.)
- The combining of elliptical and non-elliptical structure questions can bias results and so should be avoided where possible.
- Until further investigations have been carried out and firmer evidence is available, caution should be exercised in the use of negatively-phrased attitudinal items.
- The implications of including or excluding filter questions on response distributions should be considered.
- Researchers should be aware of the difficulties inherent in interpreting responses to survey questions that involve prestige names and avoid their use wherever possible.

Question sequencing

- Researchers should be aware of the potential for question order effects in self-completion questionnaires as well as in interview surveys and follow suggestions about questionnaire design accordingly.
- General questions should precede specific ones. (Recommendation based on evidence from primary research studies and expert opinion.)
- Evidence from primary research studies suggests that the "buffering" of questions is unlikely to eliminate context effects, so it is important to adhere to common survey practice of blocking questions by topic. (Recommendation based on evidence from primary research studies and expert opinion.)
- Where there is evidence that respondents may have stronger opinions on some survey topics than on others, the priority of their concerns should be determined and the survey instrument assembled to reflect them.
- Demographic questions should be placed at the end of the questionnaire. (Recommendation based on limited evidence from primary research studies combined with expert opinion.)
- Given the current lack of evidence of any ordering effects, the ordering of generic and

disease-specific measures should follow the rules for general versus specific questions. (Recommendation based on limited evidence from primary research studies combined with expert opinion.)

Response format

- The middle response category in attitude/opinion questions does not necessarily represent a position of neutrality, so it should be included.
- For factual questions, the "don't know" response may reasonably be omitted.
- If a remote scale format is used in self-completion questionnaires, the stem question should be repeated every three or four questions.
- An open space for free comment should be included in self-completion questionnaires.

Recommendations derived solely from theories of cognition and response formulation and/or expert opinion

Careful piloting of questions and their associated response categories is strongly advised, particularly when the questions have been developed especially for that survey, or when questions or scales used in a different setting or with a different population are to be used. Context gives meaning to questions and question ordering effects are rife, so questions should be piloted in context rather than in isolation. Cognitive interviewing techniques[134–137] are useful in gaining an understanding of how respondents understand and interpret questions, and of the thought processes (e.g. episodic enumeration) and heuristics (e.g. generalising from the most recently retrieved memory) they employ in responding (appendix 1).

Question wording

- The general principles of questionnaire wording (*Box 2*) should be maintained.
- The question stem and associated response categories combine to convey meaning and they should not be designed in isolation from each other.

Question ordering

- In situations where investigators are uncertain about the impact of question order on results, the order should be randomised.
- In longitudinal studies, or in those being carried out in multiple settings, the same question ordering should be maintained over time and across locations.

Response formats

- Response categories for closed questions should be mutually exclusive (i.e. unambiguous, not overlapping) and collectively exhaustive (all contingencies catered for, if necessary by the inclusion of an option of "Other, please specify").
- It should be noted that the nature of the response categories gives a subtle message about the range of ideas/concepts that the respondent should be thinking about.
- The evidence is inconsistent, so it may be preferable to label all response categories rather than only the end-points.

Recommendations for future research

Although some aspects of expert opinion concerning question wording, question ordering and the construction of response categories have not been subjected to experimental manipulation, their sense is self-evident and further investigation is unlikely to be fruitful. For example, there is no reason to believe that experts' recommendations about avoiding ambiguity in question wording would be refuted through comparisons of ambiguous and unambiguous questions.

Some other aspects of question wording, question ordering and the construction of response categories are, however, ripe for further investigation; priorities are set out below. For the most part, the authors recommend prioritising those aspects of question

BOX 2 Principles of question wording (after Moser and Kalton,[5] and Oppenheim[13])

- Use simple language
- Avoid acronyms, abbreviations, jargon and technical terms (this includes medical terms in questionnaires targeted at the general public, patients and consumers)
- Keep the question short (i.e. sentence of less than 20 words approximately)
- Avoid questions that are insufficiently specific
- Avoid ambiguity
- Avoid vague words and those with more than one meaning (e.g. "dinner")
- Avoid double-barrelled questions (i.e. those with an "and" or an "or" in the wording)
- Avoid double negatives (e.g. a negative statement followed by a "disagree" response)
- Avoid proverbs and clichés when measuring attitudes
- Avoid leading questions (e.g. "Do you agree that the NHS is under-funded?")
- Beware of loaded words and concepts
- Beware of presuming questions
- Be cautious in the use of hypothetical questions
- Do not overtax respondents' memories (e.g. by asking for detailed recall of trivial issues)

and response construction that have not been extensively studied to date. However, we also suggest that it will be important to test whether effects that have already been demonstrated in one context and with one mode of survey administration are also found in other settings and with other modes of administration, and to replicate new investigations across different modes of administration. In particular, comparisons between interviewer-administered and self-completion approaches are warranted because theories of response formulation, previous research and expert opinion suggest that different types of response bias may occur under these two modes.[52,84,117,121]

Study designs in which respondents are allocated randomly to different versions of a questionnaire (e.g. 5- versus 7-point response scales) would be appropriate in examining the effects of question wording, ordering and response category construction. However, split-half designs, in which each questionnaire contains a mix (again, randomly assigned) of items could also be considered.

As well as quantitative experimental research, qualitative methods (in particular, cognitive testing techniques[134–137]) would be appropriate in assessing how respondents comprehend questions and formulate their responses (appendix 1).

Key measures for research into question construction

In comparisons of aspects of question construction, one key measure will be the validity of the responses; in other words, is the question truly measuring what it purports to measure? Another key indicator will be the reliability of responses: is the question or questionnaire measuring things in a consistent or reproducible way? The assessment of validity and reliability is discussed in greater detail in appendix 1. These topics are also discussed in a number of key texts and articles (e.g.[22,138,139]). In addition to validity and reliability, the precision and discriminatory power of questions and their associated response categories need to be considered. Questions to which the vast majority of respondents choose the same response category are unlikely to be discriminating.[22] An examination of the distribution of responses across the response categories, using measures of spread and skewness, is advisable.

Priorities for research

Further research is required on all three main areas covered by this review. Within each area, recommendations for research are presented in priority order below.

Question wording

- Questions on the frequency and periodicity of behaviour are the key to many health-related surveys, so further research into the time-framing of questions (e.g. "1 month" versus "3 months") and of different quantifiers for time-related questions (e.g. "how many times" versus "how often") is indicated. There may be trade-offs between validity, reliability and the discriminatory power of the different quantifiers, and it will be important to take account of this in analysing data from such investigations.
- Studies of aided-recall techniques (e.g. bounded recall) for memory questions are recommended. No research on this topic was identified.
- Comparative studies should be carried out of the different methods suggested by Sudman and Bradburn[7] (described in the first part of this chapter) for deliberately loading threatening or sensitive questions in order to obtain more valid responses. No research on this topic was identified.
- Conventional wisdom suggests that a mix of positively and negatively worded statements should be used in measuring attitudes, but the limited evidence from the one study identified on this topic[89] concluded that the inclusion of negative items in attitudinal questionnaires may impair rather than increase the validity of survey results. Further research into the impact of mixing positive and negative statements is therefore recommended.
- There was limited evidence on the impact of filter questions. The authors therefore advocate comparisons of the inclusion and exclusion of filter questions and suggest that these should focus on: filtering out respondents with no preformulated opinions before asking detailed questions about attitudes; using filter questions to avoid asking detailed questions of people who have no knowledge of a topic; and filtering out those respondents who have never engaged in a particular form of behaviour.

Question sequencing

- Theories of respondent behaviour suggest that question ordering effects may be reduced in self-completion questionnaires (because the respondent has the opportunity to preview all the questions before responding), but empirical evidence on this topic is limited. The authors therefore advocate that research into the effect of question ordering should concentrate on self-completion questionnaires. Theories of respondent behaviour suggest that ordering effects are most marked in respect of attitudinal questions[115,116] and the authors recommend

that these should be the first priority in future investigations.
- Social desirability bias may occur when behaviour questions are asked after knowledge questions on a related topic (e.g. questions on personal dietary behaviour after items on knowledge of good eating practice), so comparisons of the relative positions of these sets of questions are warranted. No existing studies on this specific aspect of question ordering were identified.
- The apparent relevance and "ease of answering" of opening questions may influence the decision to respond,[1,140] so comparisons of more and less salient opening items are indicated.

Response categories

- The ordering of response categories may lead to response bias of both recency and primacy effects.[52,84] Further comparative studies of alternative ordering are therefore desirable. This is particularly true for questions on sensitive topics, where it has been suggested that the categories should be ordered from the least to the most socially desirable.[7]
- Recency effects appear to be more common in interviewer-administered surveys.[52,84] The authors recommend research into ways of minimising such effects (e.g. the use of prompt cards; whether multiple-step approaches are any more effective than single-step methods; what techniques can be used in telephone surveys).
- Sudman and Bradburn[7] have suggested that analogue scales (e.g. ladders, clocks, thermometers) may be effective for numerical scales that have many points. No studies of such approaches were located and the authors recommend that such analogue methods should be compared with more conventional numerical scales.
- It has been suggested that increased precision may be achieved through the use of seven rather than five response categories,[36] especially in Likert-type scales, and there is some evidence for this.[141] There is little evidence, however, for further enhancement of precision beyond seven categories. Further research into the reliability and discriminatory power of 5- versus 7-point (or more finely graded) scales is recommended.
- Findings from identified comparative studies of the labelling of all scale points compared with attaching verbal descriptors to end-points only are equivocal.[125,130] Further research into this topic is therefore desirable.

TABLE 6 *Question wording*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Criteria for comparison	Main findings
Markum, 1976[87]	RCT	4	Health: women's experiences of menstrual symptoms	University students and employees (USA)	Self-completion survey	***Neutral vs. non-neutral questions:*** Masked MMDQ (47) Unmasked MMDQ (47) In the masked version, respondents were not told that the listed symptoms were menstrual	Reporting of symptoms (mean no. reported)	Mean scores (SDs) for respondents in all stages of menstrual cycle: Masked 98.7 (24.8); unmasked 95.4 (22.6) Mean difference = –3.3 (95% CI, –13.0 to 6.4) No significant differences in mean scores on controlling for stage in the menstrual cycle 70% of respondents in masked group reported having no idea about what the questionnaire was assessing
Larsen *et al.*, 1987[95]	RCT (study 1) Systematic allocation (study 2)	4 (study 1) 3.5 (study 2)	Health: frequency of headache + tooth-brushing	University students (study 1) + members of general public (study 2) (USA)	Self-completion survey	***Time-framing in question wording:*** Study 1: 60 male, 60 female university students (120) "Frequently" quantifier: 30 male, 30 female (60) "Occasionally" quantifier: 30 male, 30 female (60) Study 2: General public (355) 4 × 2 versions (defined by headache and tooth-brushing questions) of questionnaire with systematic allocation to each 4 versions of question on headache: HV1 – "Do you get headaches **frequently**? If so, how often?" (86) HV2 – "Do you get headaches **occasionally**? If so, how often?" (91) HV3 – "How **frequently** do you get headaches?" (88) HV4 – "How **often** you get headaches?" (90) 2 versions of question on tooth-brushing: TV1 – "How **frequently** do you brush your teeth?" TV2 – "How **often** do you brush your teeth?"	Reporting of headaches (whether reported; mean no. reported) Reporting of tooth-brushing (mean no. sessions reported)	Study 1: Mean no. headaches/week: "Frequently" quantifier = 1.15; "occasionally" quantifier = 0.71 Difference reported as ns; insufficient data to calculate 95% CI Reporting not having headaches at all: "Frequently" quantifier = 57%; "occasionally" quantifier = 38% RR = 0.68 (95% CI, 0.46 to 1.00) "occasionally" vs. "frequently" Study 2: Reporting headaches: HV1 10%; HV2 54%; HV3 38%; HV4 46% RR = 5.15 (95% CI, 2.69 to 9.82) HV2 vs. HV1 RR = 3.58 (95% CI, 1.83 to 7.03) HV3 vs. HV1 RR = 4.35 (95% CI, 2.25 to 8.41) HV4 vs. HV1 RR = 0.70 (95% CI, 0.50 to 0.97) HV3 vs. HV2 RR = 0.85 (95% CI, 0.63 to 1.14) HV4 vs. HV2 RR = 1.21 (95% CI, 0.85 to 1.73) HV4 vs. HV3 Mean no. headaches/week (1st figure – only those indicating that they did get headaches; 2nd figure – all respondents, allocating a value of 0 to those reporting none): HV1 4.1 (0.42) HV2 1.5 (0.82) HV3 1.6 (0.60) HV4 1.2 (0.57) Difference for all respondents reported as ns; insufficient data to calculate 95% CI Mean no. tooth-brushing sessions/day: TV1 1.83; TV2 1.88 Difference reported as ns; insufficient data to calculate 95% CI

continued

TABLE 6 contd *Question wording*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Criteria for comparison	Main findings
Salvendy, 1976[88]	RCT	3	Non-health: attitudes to religion and transport	College professors + city dwellers (all male) (USA)	Postal survey	***Direct vs. indirect questions:*** Direct question version 500 per topic (1000) Indirect question version 500 per topic (1000)	Instrument response rates	Overall response rates: no significant difference ($p > 0.75$) with respect to style of question (direct vs. indirect) or topic. Actual values not reported in numerical form; insufficient data to calculate RR and 95% CI Response rate to direct version significantly higher than to indirect version among general public, but not among college professors ($p < 0.005$); actual values not reported
Schriesheim and Hill, 1981[89]	RCT	4	Non-health: views on manager behaviour (using written description of fictitious manager)	Business students (USA)	Self-completion survey	***Positive vs. negative questions:*** Positively worded items (50) Negatively worded items (50) Mixed wording (50) Used Leader Behaviour Description Questionnaire – XII, with items reworded appropriately	Accuracy of response (with respect to researchers' assessment of written description)	Mean scores (SDs) across all items – lower scores denote greater accuracy: Positive 7.1 (4.7); negative 10.4 (7.6); mixed 10.5 (6.1) Mean difference = 3.4 (95% CI, 1.2 to 5.6) positive vs. negative Mean difference = 3.3 (95% CI, 0.8 to 5.8) positive vs. mixed Mean difference = 0.1 (95% CI, –2.7 to 2.8) negative vs. mixed Decrement in accuracy appeared to be a function of the negatively worded items and not of their contextual effect on positive items There was a small detrimental context effect for negatively worded items when mixed with positively worded items
Bishop *et al.*, 1982[90]	RCT	4	Non-health: views on public affairs issues	Householders with telephones (USA)	Telephone survey	***Neutral vs. non-neutral questions:*** P1: one side of issue presented (agree/disagree format) P2: both sides of issue presented (substantive choice format) ***One- vs. two-sided attitudinal questions:*** FQ1: "... Do you have an opinion on this or not?" FQ2: "... Have you been interested enough in this to favour one side over the other?" Total sample size of 2296 across 2 replications of a 2×2 factorial design (i.e. approximately 1148/form)	Expression of opinion (% expressing an opinion; % agreeing with single-sided statement)	% respondents expressing an opinion when presented with substantive choice vs. when simply asked to agree/disagree: Difference significant ($p < 0.05$) for 2/5 comparisons with FQ1 Difference approaches significance ($p < 0.10$) for 1/4 comparisons with FQ2 In all cases, higher %s for those offered a substantive choice (Because of multiple comparisons, space constraints preclude presentation of RRs and CIs) % respondents expressing an opinion did not differ significantly with form of filter question % respondents endorsing initial single-sided statement when presented with substantive choice vs. when simply asked to agree/disagree: Difference significant ($p < 0.05$) for 3/5 comparisons with FQ1 Difference significant ($p < 0.01$) for 2/4 comparisons with FQ2 In all cases, lower %s for those offered a substantive choice (Because of multiple comparisons, space constraints preclude presentation of RRs and CIs) Less well-educated informants were more likely to acquiesce to one-sided agree/disagree forms.

continued

TABLE 6 contd *Question wording*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Criteria for comparison	Main findings
Bishop *et al.*, 1983[91]	RCT	3	Non-health: views on public affairs issues	Householders with telephones (USA)	Telephone survey	***Wording of filter questions:*** Standard form: no filter question Filter A: "Do you have an opinion on this matter?" Filter B: "Have you been interested enough in this to favour one side over the other?" Filter C: "Have you thought much about this issue?" Filter D: "Where do you stand on this issue, or haven't you thought much about it?" Filter E: "Have you already heard or read enough about it to have an opinion?" Sample size varied across 5 constituent studies for different issues in each study No. and nature of forms used varied from item to item within each survey Studies 1 and 2: standard form + filters A–D Study 3: standard form + filter A Studies 4 and 5: standard form + filters A, C and E	Provision of DK response (distribution of DK responses)	Overall, the increment in DK responses due to adding a filter ranged from 4.5% to 46.9% (mean 22%); effect of filter depended on both the content of the item and the wording of the filter % DK responses: Studies 1 and 2 combined: significant differences ($p < 0.05$) in distribution of DK responses across various filters for 3/8 substantive questions (2 others approach statistical significance: $0.10 < p < 0.05$); filters B, C and D consistently screened out more respondents than filter A Studies 4 and 5 combined: significant differences ($p < 0.05$) in distribution of DK responses across various filtered forms for 4/7 substantive questions; filter C consistently screened out more respondents than filter A; filter E represented a stronger filter than filter C (Because of multiple comparisons, space constraints preclude presentation of RRs and CIs) For roughly half the topics studied, the inclusion of a filter question created a statistically significant difference in the distribution of substantive responses

continued

TABLE 6 contd *Question wording*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Criteria for comparison	Main findings
Bishop *et al.*, 1986[92]	RCT	4	Non-health: views on public affairs issues (including fictitious issues)	Householders with telephones (USA)	Telephone survey	***Wording of filter questions:*** Form A: filter allowing respondents to indicate they had not thought about issue (397–408) Form B: no filter, but respondents allowed to give DK response (395–411) Form C: no filter, respondents pressed to select substantive response (397–404) 2 replications, with 3 fictitious topics/replication: different achieved sample sizes/ replication and topic (Also manipulated: ordering of fictitious and genuine issues; response categories (inclusion of DK category)) See also *Table 8*	Expression of opinion (% expressing an opinion)	% respondents offering an opinion on fictitious issues significantly different across all three forms; % always lower for filtered form A ($p < 0.0001$ for all 6 comparisons) (Because of multiple comparisons, space constraints preclude presentation of RRs and CIs) % respondents offering an opinion on fictitious issues significantly higher when DK response probed (i.e. form C) ($p < 0.05$ for all comparisons) (Because of multiple comparisons, space constraints preclude presentation of RRs and CIs) The less knowledgeable people were about an issue, the more easily they could be pressurised into giving an opinion about it
Schuman *et al.*, 1986[93]	RCT	4	Non-health: views on public affairs issues (including threat of nuclear war)	Householders with telephones (USA)	Telephone survey	***Open-ended vs. closed questions:*** Open-ended (174–282) Closed (170–261) 5 replications with different numbers of respondents/ replication	Range of responses given	Open-ended questions produce more "non-common" category responses (Because of multiple comparisons, space constraints preclude presentation of RRs and CIs) For 4/5 replications, the rankings (based on % of respondents mentioning that category) of 4 "common" response categories were identical between open-ended and closed question forms (Because of multiple comparisons, space constraints preclude presentation of RRs and CIs) The less frequently an issue was mentioned spontaneously in response to an open-ended question, the greater its increase in choice at a closed version of the same question

continued

TABLE 6 contd *Question wording*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Criteria for comparison	Main findings
Blair and Burton, 1987[94]	RCT	4	Non-health: frequency of undertaking various high-participation activities	Householders with telephones (USA)	Telephone survey	***Time-framing in question wording:*** 2 weeks (128) [108] 2 months (128) [108] 6 months (128) [116] How many (192) [169] How often (192) [163] 64/cell randomised in a 3 × 2 factorial design [Nos of respondents included in analysis of cognitive processes]	Reported cognitive processes employed (including episodic enumeration)	Respondents reported 12 distinct cognitive processes for estimating behavioural frequency Using episodic enumeration: 2 weeks 56%; 2 months 25%; 6 months 4% How many 30%; how often 26% RR = 0.45 (95% CI, 0.31 to 0.65) 2 months vs. 2 weeks RR = 0.08 (95% CI, 0.03 to 0.19) 6 months vs. 2 weeks RR = 0.17 (95% CI, 0.07 to 0.43) 6 months vs. 2 months RR = 0.85 (95% CI, 0.60 to 1.21) how often vs. how many No significant interaction between time-frame and question wording
Barnes and Dotson, 1989[96]	RCT	3.5	Non-health: attitudes to a range of social issues	TV viewers (USA)	Interview survey	***Elliptical vs. non-elliptical questions:*** Elliptical version first (401) Non-elliptical version first (398) Total 825 randomised; achieved sample size 799 Ordering of topic area of preceding question also varied	Nature of response (mean score on response scale)	Mean response across all 15 question pairs (on 5-point scale: score of <3 indicates agreement; >3 disagreement): Elliptical version first 3.42; non-elliptical version first 3.34 Mean difference = 0.08 (95% CI, –0.06 to 0.22) Elliptical structure produced more disagreement for 5 questions; more agreement for 8 questions; a shift towards the neutral position for 2 questions Differences were reported to be significant ($p < 0.05$) for 4/15 questions (Insufficient data presented to calculate 95% CIs)

continued

TABLE 6 contd *Question wording*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Criteria for comparison	Main findings
Smith and Squire, 1990[97]	RCT (initial random sample with sequential allocation to groups)	4	Non-health: views on judicial elections (study 1) and tax indexing (study 2)	Voters (USA)	Not specified	***Use of prestige names:*** Study 1: Prestige name not included (408) Prestige name included (402) Study 2: Prestige name not included (603) Prestige name included (614)	Expression of opinion (% expressing an opinion)	Study 1: Expressing a negative opinion for each of 4 items: Prestige name not included: 11%; 7%; 8%; 5% Prestige name included: 27%; 21%; 22%; 15% RR = 2.46 (95% CI, 1.79 to 3.38) included vs. not included (item 1) RR = 2.94 (95% CI, 1.97 to 4.38) included vs. not included (item 2) RR = 2.71 (95% CI, 1.86 to 3.94) included vs. not included (item 3) RR = 3.04 (95% CI, 1.87 to 4.95) included vs. not included (item 4) Inclusion of prestige names nullified trend towards greater propensity of more highly educated to express a definite opinion Study 2: Expressing no opinion on each of 2 items: Prestige name not included: 26%; 20% Prestige name included: 25%; 15% RR = 0.95 (95% CI, 0.62 to 1.47) included vs. not included (item 1) RR = 0.75 (95% CI, 0.58 to 0.96) included vs. not included (item 2) Using prestige names added a partisan component to responses; this was most marked when respondents knew less about the topic

DK, "don't know" response; MMDQ, Moos Menstrual Distress Questionnaire

TABLE 7 *Question sequencing*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Criteria for comparison	Main findings
Schuman *et al.*, 1981[107]	RCT	4	Health: views on various social issues (including abortion)	Householders with telephones (USA)	Telephone survey	***Ordering of general vs. specific questions:*** Specific question first (440) General question first (440) (Achieved samples 293 and 305 respectively)	Expression of opinion (% expressing agreement)	Agreement with general question: Specific first 48%; general first 61% RR = 1.26 (95% CI, 1.09 to 1.46) general first vs. specific first Agreement with specific question: Specific first 84%; general first 83% RR = 0.99 (95% CI, 0.92 to 1.06) general first vs. specific first Order effect not affected by religion, gender or educational level
Colasanto *et al.*, 1992[108]	RCT	4	Health: public knowledge about AIDS transmission	Householders with telephones (USA)	Telephone survey	***General context effects:*** Question on blood transfusion before question on blood donation (505) Question on blood transfusion after question on blood donation (505)	Expression of opinion (% expressing agreement)	Agreeing that AIDS can be transmitted by blood donation: Question on transfusion before that on donation 43% Question on transfusion after that on donation 52% RR = 1.20 (95% CI, 1.06 to 1.37) transfusion question after donation question vs. transfusion question before donation question
Barry *et al.*, 1996[109]	RCT (initial random sample with sequential allocation to groups)	4	Health: generic and condition-specific health status	Men with benign prostatic hyperplasia (USA)	Postal survey	***Ordering of generic vs. condition-specific instruments:*** Condition-specific first (198) Generic first (194)	Distribution of scores on constituent scales of health status/quality of life instruments	No statistically significant differences in the distribution of scores between the groups for scores on any of the 8 generic health status scales (95% CIs for differences in means all include zero) No statistically significant differences in the distribution of scores between the groups for scores on any of the three condition-specific scales (95% CIs for differences in means all include zero) No significant differences between the groups in magnitude of correlation between scores on the generic and condition-specific scales

continued

TABLE 7 contd *Question sequencing*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Criteria for comparison	Main findings
Jones and Lang, 1980[110]	RCT	4	Non-health: details of house purchase	House purchasers (USA)	Postal survey	***General context effects:*** Order 1 (1463) (Demographic questions – semantic differential attributes scales – anchored similarity judgement scales) Order 2 (1463) (Demographic questions – anchored similarity judgement scales – semantic differential attributes scales) Total sample size of 2926 in a $2 \times 2 \times 3 \times 2$ factorial design (122 per group) (Also manipulated: no. and timing of contacts; covering letter (nature of appeal); sponsorship) See also *Tables 16, 22* and *23*	Instrument response rates Sample composition bias (difference in distribution of variables between entire sampling frame and respondents) Response bias (difference between reported and actual purchase prices and dates)	Instrument response rates: order 1, 28%; order 2, 23% RR = 0.82 (95% CI, 0.72 to 0.93) order 2 vs. order 1 Sample composition bias: For distribution of house purchase prices, sample composition bias found for order 1. This format yielded an invalid, positively biased estimate of distribution of prices; no such bias was found for order 2 For distribution of house purchase dates, no sample composition bias found for either order Response bias: No significant main effect for order on response bias with respect to house purchase price No significant main effect for order on response bias with respect to house purchase date Significant notification × order interaction; order 1 led to negative bias (i.e. purchase date being recalled as earlier than it really was) only in group who were prenotified
McFarland, 1981[111]	RCT (initial random sample with sequential allocation to groups)	4	Non-health: views about a range of social issues	Householders with telephones (USA)	Telephone survey	***Ordering of general vs. specific questions:*** Specific first (258) [244–249] General first (258) [247–252] 258 randomised to each order [No. usable responses per item]	Expression of opinion (% expressing an opinion)	Agreeing that energy problem is "extremely serious": Specific first 39%; general first 34% RR = 0.88 (95% CI, 0.70 to 1.11) general first vs. specific first Agreeing that economy "will get better": Specific first 17%; general first 13% RR = 0.75 (95% CI, 0.49 to 1.15) general first vs. specific first Stating themselves to be "very interested" in politics: Specific first 42%; general first 28% RR = 0.67 (95% CI, 0.52 to 0.85) general first vs. specific first Stating themselves to be "very interested" in religion: Specific first 64%; general first 56% RR = 0.88 (95% CI, 0.76 to 1.01) general first vs. specific first Order had little effect on strength of relationship between general and specific; only 2/17 relationships significant at 5% level Order effects particularly marked for "interest" questions; effects were consistent across gender and levels of education

continued

TABLE 7 contd *Question sequencing*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Criteria for comparison	Main findings
Sigelman, 1981[112]	RCT	4	Non-health: views of US presidency	Householders with telephones (USA)	Telephone survey	***General context effects:*** Question on presidential popularity before questions on range of social issues (373) Question on presidential popularity after questions on range of social issues (373) Most of the questions on social issues were "politically charged" in a negative direction	Expression of opinion (% expressing an opinion)	Respondents willing to express an evaluation of president: Presidential popularity question first 80% Presidential popularity question after social issues questions 89% RR = 1.10 (95% CI, 1.03 to 1.18) popularity question after social issues question vs. popularity question first Respondents expressing approval of president: Presidential popularity question first 53% Presidential popularity question after social issues questions 52% RR = 0.98 (95% CI, 0.83 to 1.14) popularity question after social issues question vs. popularity question first
Schuman *et al.* 1983[113]	RCT	4	Non-health: views about USA–USSR relations	Householders with telephones (USA)	Telephone survey	***Contiguous vs. non-contiguous questions:*** Version A: question about a Communist reporter first, followed immediately by one about an American reporter (117) Version B: order of version A items reversed (107) Version C: question about the American reporter was first and was separated by 17 items from the question about the Communist reporter (107)	Expression of opinion (% expressing agreement)	Answering "yes" to item regarding freedom of Communist reporters to report from USA: Version A 44%; version B 70%; version C 66% RR = 0.63 (95% CI, 0.50 to 0.80) version A vs. version B RR = 0.67 (95% CI, 0.53 to 0.85) version A vs. version C RR = 1.06 (95% CI, 0.88 to 1.27) version B vs. version C Context effects not diminished by separating key items within the same questionnaire
Spector and Michaels, 1983[114]	RCT	3	Non-health: attitudes to work and job satisfaction	Employees of mental health centre (USA)	Postal survey	***General context effects:*** Satisfaction questions before perception questions (90) Satisfaction questions after perception questions (90) (Achieved sample sizes 55 and 69 respectively)	Correlations between responses to questions	Of 300 possible comparisons across versions of correlations between all pairs of the 25 variables in the questionnaire, only 13 (4.3%) were significantly different ($p < 0.05$). The few significant differences were within the range expected by chance under the null hypothesis (Insufficient data to calculate 95% CIs for correlations)

continued

TABLE 7 contd *Question sequencing*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Criteria for comparison	Main findings
Tourangeau *et al.*, 1989[115]	RCT	3.5	Non-health: views on public affairs issues (abortion, welfare, Nicaragua, Star Wars)	Householders with telephones (USA)	Telephone survey	***General context effects:*** Side of context items preceding target item (liberal or traditional) Mode of presentation of context items (blocked or scattered) Level of agreement with context items (high or low) Order of issues Total sample size of 2556 in a $2 \times 2 \times 2 \times 2$ factorial design addressing 4 issues (achieved sample of 1251 for first interview; 1056 for second)	Consistency of response (count of responses to target items that were consistent with responses to related context items) Expression of opinion (% expressing opinion)	Target responses consistent with context: Mean 2.27 for liberal set first; mean 1.94 for traditional set first ($p < 0.001$; insufficient data to calculate 95% CI) % expressing particular opinions (e.g. favouring legalised abortion): For all 4 issues, context effect was in the direction of "side" of context issues but size of effect varied across issues; 95% CIs for the RR for liberal vs. traditional did not include unity for welfare and Nicaragua items Blocking had a significant main effect for the Nicaragua and Star Wars items and there was a mode × side interaction for the Star Wars item (Because of multiple comparisons, space constraints preclude presentation of RRs and 95% CIs) "Carry-over" effects were related to 2 factors: a close target-context relationship and contiguity in the questionnaire; attitude conflict and centrality were also important
Tourangeau *et al.*, 1989[116]	RCT	4	Non-health: views on public affairs issues	Householders with telephones (preference given to males) (USA)	Telephone survey	***Contiguous vs. non-contiguous questions:*** 10 groups (approximately 114 in each) 10 different versions of a questionnaire in which the order of 6 target questions and the order and content of context items were varied	Distribution of responses	For 5/6 target issues, responses to the target questions varied significantly across the groups ($p < 0.001$). Results indicate that context effects are not an isolated occurrence (Because of multiple comparisons, space constraints preclude presentation of RRs and 95% CIs) For 4/6 target issues, context effects were larger when respondents' beliefs about the target issue were both mixed and important to them; for 2/6, this effect was statistically significant ($p < 0.05$ and $p < 0.005$ respectively) (Because of multiple comparisons, space constraints preclude presentation of RRs and 95% CIs)

continued

TABLE 7 contd *Question sequencing*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Criteria for comparison	Main findings
Ayidiya and McClendon, 1990[117]	RCT	4	Non-health: views on a range of social issues	Householders with telephones (USA)	Postal survey	***General context effects: Ordering of general vs. specific questions:*** Version 1 (266) Version 2 (266) Questionnaire versions differed with respect to: order in which questions were presented; order of response categories; inclusion of "no opinion" filter questions; provision of middle alternative response categories; agree–disagree vs. forced choice options (Also manipulated: response categories – provision of "middle" category) See also *Table 8*	Distribution of responses	Significant question order effects ($p < 0.05$) for 1/4 comparisons; another approached statistical significance ($p = 0.07$) Tendency towards primacy effect ($p = 0.08$) with respect to response categories for 1/3 comparisons No evidence of recency effects (p = ns) with respect to response categories for any of 3 comparisons % answering DK significantly higher ($p < 0.01$) with inclusion of "no opinion" filter for all 4 items considered % choosing "middle alternative" significantly greater ($p < 0.01$) when this option offered explicitly % agreeing with statement greater for "agree–disagree" format than for forced-choice format; difference significant for 2/3 items ($p = 0.03$ and $p < 0.01$ respectively) (Because of multiple comparisons, space constraints preclude presentation of RRs and 95% CIs)
Roberson and Sundstrom, 1990[118]	RCT	3.5	Non-health: attitudes towards aspects of employment conditions	Office employees (USA)	Self-completion survey	***General context effects:*** 99 per each of 6 × 2 groups Order of topics: 1 prioritised by employees + 5 randomised Position of demographic questions: first or last within questionnaire	Instrument response rates Attitude scores (global; topic-specific)	***Instrument response rates:*** Ordering of topics: Average across all randomised ordering: 78% Order prioritised by employees: 96% (All random orders had lower response rates than prioritised order) RR = 1.23 (95% CI, 1.18 to 1.29) prioritised vs. average for all random Ordering of demographic items: Demographic questions first 77%; demographic questions last 85% RR = 1.11 (95% CI, 1.05 to 1.17) last vs. first ***Global and topic-specific attitude scores:*** Significantly higher ($p < 0.05$) global attitude scores (more favourable) for order prioritised by employees Topic attitude scores significantly higher ($p < 0.05$) for 3/6 topics for order prioritised by employees (Insufficient data to calculate mean differences and 95% CIs)

continued

TABLE 7 contd *Question sequencing*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Criteria for comparison	Main findings
Tenvergert *et al.*, 1992[119]	Historically controlled study	4	Health: public attitudes to legal abortion	Adults on electoral register (Canada)	Interview survey	***Ordering of general vs. specific questions:*** 1984: order of questions DEFECT/NOMORE/HEALTH 560 (effective sample size 395) 1987: order of questions HEALTH/NOMORE/DEFECT 620 (effective sample size 375) 1988: order of questions DEFECT/NOMORE/HEALTH 584 (effective sample size 390) NOMORE: approval of abortion if married woman wants no more children (general) DEFECT: approval of abortion if strong chance of defective baby (specific) HEALTH: approval of abortion if woman's health seriously endangered (specific)	Expression of opinion (% expressing agreement)	Positively endorsing NOMORE: 1984 43%; 1987 56%; 1988 52% RR = 1.29 (95% CI, 1.11 to 2.19) 1987 vs. 1984 RR = 1.20 (95% CI, 1.03 to 1.86) 1988 vs. 1984 RR = 0.93 (95% CI, 0.81 to 1.13) 1988 vs. 1987 Positively endorsing DEFECT: 1984 86%; 1987 94%; 1988 86% RR = 1.09 (95% CI, 1.04 to 1.15) 1987 vs. 1984 RR = 1.00 (95% CI, 0.95 to 1.06) 1988 vs. 1984 RR = 0.91 (95% CI, 0.87 to 0.96) 1988 vs. 1987 Positively endorsing HEALTH: 1984 90%; 1987 94%; 1988 95% RR = 1.04 (95% CI, 0.99 to 1.08) 1987 vs. 1984 RR = 1.05 (95% CI, 1.01 to 1.09) 1988 vs. 1984 RR = 1.01 (95% CI, 0.98 to 1.05) 1988 vs. 1987
Serdula *et al.*, 1995[120]	Historically controlled study	4	Health: surveillance of behavioural risk factors	Householders with telephones (USA)	Telephone survey	***General context effects:*** Question order 1 (117,827) Question order 2 (114,025) Question order 1, used in 1985–1988, had questions on self-reported height and weight immediately before question on attempted weight loss; question order 2, used in 1989, 1991 and 1992, had questions on height and weight at end of questionnaire	Reported attempted weight loss	Reporting attempted weight loss: Question order 1: 29% (males); 48% (females) Question order 2: 26% (males); 41% (females) RR = 0.90 (95% CI, 0.88 to 0.91) order 2 vs. order 1 (males) RR = 0.85 (95% CI, 0.84 to 0.87) order 2 vs. order 1 (females)

TABLE 8 *Response categories*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Criteria for comparison	Main findings
Miller, 1984[82]	RCT	2	Health: family health and life satisfaction	Householders with telephones (USA)	Telephone survey	***Ordering of response categories:*** 1-step approach (2150) 2-step approach (2150)	Mean question scores Distribution of responses	Mean score across 5 health satisfaction questions: 1-step approach 5.13; 2-step approach 5.97 (significance not stated; insufficient data to calculate CI) Slight tendency for 2-step version to produce higher means, but difference significant (5.77 vs. 5.97) only for 1/5 items (insufficient data to calculate CI) Distribution of responses across 7 points of response scale: Significantly different between the 2 versions ($p < 0.001$ in all cases; insufficient data to calculate CIs) Effect of question form was not affected by perceived salience of health (i.e. how often respondent thought about health) or by whether the respondent suffered from a chronic condition
Poe *et al.*, 1988[123]	RCT	4	Health: healthcare in last year of life	Relatives of recently deceased persons (USA)	Postal survey	***Inclusion of DK response category:*** DK box (678) No DK box (682) In the version with explicit DK boxes, respondents were told to mark the box if one was provided or put a ? in the answer space if they did not know the answer; in the version without DK boxes, respondents were told to put a ? in the answer space; all explicit DK responses and those where a ? was placed in the answer space were coded as DK for purposes of analysis	Instrument response rates Item non-response rates	Instrument response rates: DK boxes 62%; no DK boxes 58% RR = 1.04 (95% CI, 0.96 to 1.14) no DK boxes vs. DK boxes Average items marked DK or with a ?: DK boxes 7%; no DK boxes 2% RR = 3.99 (95% CI, 1.77 to 9.00) no DK boxes vs. DK boxes Average items left blank: DK boxes 11%; no DK boxes 13% RR = 0.85 (95% CI, 0.59 to 1.24) no DK boxes vs. DK boxes Version without DK boxes had average substantive responses 3.2% higher than the version with DK boxes (difference ns) For many specific items the % of substantive responses was appreciably higher for the version without DK boxes; the % of substantive responses for the version without DK boxes was lower for only 1/187 items For a quarter (46/187) of the specific items, the % of substantive responses was significantly higher for the version without DK boxes; differences of ≥25% were observed for 4/187 items; differences ≥5% were observed for 51 items (of which 40 were statistically significant) For the majority of items (89%) there was no significant difference in the distribution of substantive responses between the 2 versions (Because of multiple comparisons, space constraints preclude presentation of RRs and 95% CIs)

continued

TABLE 8 contd *Response categories*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Criteria for comparison	Main findings
Stem *et al.*, 1978[124]	RCT	3	Non-health: attitudes to marketing course	Business students (USA)	Self-completion survey	***Remote vs. adjacent placement of labels:*** Adjacent scale format (290) Remote scale format (282) Adjacent format presented rating scale after each question; remote format presented rating scale at top of each page (Ordering of questions also varied)	Mean question scores Distribution of responses	Remote format resulted in less skewed responses (i.e. elicited more neutral responses than the adjacent scale format; $p = 0.011$) Neutral shift was not dependent on the distance of the stimulus from the scale ($r = 0.119$) for mean scores, but there were significant increases in variance using the remote format; the further from the stimulus the greater the variance Remote format tended to produce higher mean scores for individual items but difference significant ($p < 0.05$) for only 2/10 questions Remote format tended to produce more variable scores for individual items but difference significant ($p < 0.05$) for only 1/10 questions (Insufficient data to calculate 95% CIs for mean differences)
Frisbie and Brandenburg, 1979[125]	RCT	4	Non-health: views about education	University students (USA)	Self-completion survey	***Labelling of response categories:*** A: all scale points labelled alphabetically (899) B: only end-points labelled alphabetically (902) C: all scale points labelled numerically (912) D: only end-points labelled: numerically (903) (6 of the 8 items had 5 scale points; the others had 4)	Mean question scores	Mean score on individual items: Significant differences ($p < 0.01$) for 6/8 items; in all cases, means were higher (more favourable rating) for those for whom only end-points labelled (i.e. groups B + D); all significant differences were for the 5-point scales No significant differences between groups with alphabetical labels (A + B) and those with numerical labels (C + D) (Because of multiple comparisons, space constraints preclude presentation of RRs and 95% CIs)
Edvardsson, 1980[126]	Non-random con-current controlled study (system-atic allocation)	4	Non-health: views about environmental problems	Psychology students (USA)	Self-completion survey	***Ordering of response categories:*** Positive responses first (65) Negative responses first (65) Questionnaire contained 53 questions grouped into 6 sets; no. response categories in scale varied from set to set: 4, 3, 2, 3, 4, 5 respectively	Distribution of responses	Tendency to select responses to the left of the scale, but... No significant differences in response distributions for any of the 6 categories or for any of the individual questions (Insufficient data to calculate 95% CIs)

continued

TABLE 8 contd *Response categories*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Criteria for comparison	Main findings
Hawkins and Coney, 1981[127]	RCT	4	Non-health: evaluation of a fictitious public agency	Lawyers + householders with telephones (USA)	Postal survey	***Inclusion of DK response category:*** Inclusion (≅500) Non-inclusion (≅500) Study used a 2 × 2 × 2 factorial design with groups defined by: lawyer vs. general public; inclusion vs. non-inclusion of explicit DK response category; 5 vs. 1 fictitious issue	Instrument response rates (adjusted for undeliverable questionnaires) Expression of opinion (% respondents expressing an uninformed opinion)	Instrument response rates: Inclusion of DK category 31%; non-inclusion of DK category 28% RR = 0.91 (95% CI, 0.74 to 1.11) non-inclusion of DK vs. inclusion of DK category Respondents responding to chosen question on fictitious issue (i.e. "expressing an uninformed opinion"): Inclusion of DK category 63%; non-inclusion of DK category 95% RR = 1.49 (95% CI, 1.31 to 1.70) non-inclusion of DK vs. inclusion of DK category
Israel and Taylor, 1990[128]	RCT	4	Non-health: practices and views about beef production	Beef producers (USA)	Postal survey	***Ordering of response categories:*** Form A (160) Form B (184) Order of response categories for each of 9 questions varied between the 2 forms; 5 questions required a single response, 4 required multiple responses; 2 (1 of each type) were "attribution" questions, where a social acceptability effect was anticipated; no. response categories range 3–9	Instrument response rates Distribution of responses	Instrument response rates: Form A 48%; form B 50% RR = 1.04 (95 % CI, 0.84 to 1.29) form B vs. form A Distribution of responses: On 2/9 questions, there was a significant difference between the 2 forms ($p < 0.001$ and $p = 0.02$ respectively) Order effects (i.e. differences in distribution of responses) were not found for any of the 4 single-response items that were not attributive An order effect was found for 1 of the 3 non-attributive multiple-response questions An order effect was found for the multiple-response attributive question (Because of multiple comparisons, space constraints preclude presentation of 95% CIs)
Swan and Epley, 1981[129]	RCT	4	Non-health: licensing for real estate brokerage	Real estate brokers (USA)	Postal survey	***Labelling of response categories:*** Narrow income bands ≅500) Wide income bands ≅500) Instructed to check only 1 income band ≅500) Allowed to check 2 adjacent income bands ≅500) Total sample size 1000; 2 × 2 factorial study design	Instrument response rates (return; completion)	Instrument response rates: Narrow income bands 50%; wide 50% RR = 1.00 (95% CI, 0.92 to 1.09) narrow vs. wide Checking 1 band only allowed 49%; two adjacent bands allowed 51% RR = 1.04 (95% CI, 0.95 to 1.14) 2 vs. 1 allowed Instrument completion rates (based on those returned): Narrow income bands 89%; wide 92% RR = 0.97 (95% CI, 0.93 to 1.01) narrow vs. wide Checking 1 band only allowed 95%; 2 adjacent bands allowed 86% RR = 0.90 (95% CI, 0.87 to 0.94) 2 allowed vs. 1 allowed

continued

TABLE 8 contd *Response categories*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Criteria for comparison	Main findings
Lam and Klockars, 1982[130]	RCT	4	Non-health: views about quality of academic courses and teaching	University students (USA)	Self-completion survey	***Labelling of response categories – scale anchors:*** Equally spaced (93) Positively packed (98) Negatively packed (92) Only end-points labelled (92) (Achieved sample sizes)	Mean question scores	Mean scores across 4 selected questions: Equally spaced 3.296; positively packed 3.112; negatively packed 3.454; only end-points labelled 3.299 $p < 0.05$ for comparison of all 4 groups p = ns for comparison of equally spaced and only end-points labelled $p < 0.05$ for all other pair-wise comparisons (Because of multiple comparisons, space constraints preclude presentation of mean differences and 95% CIs) Order of means: negatively packed > only end-points labelled > equally spaced > positively packed
Trice and Dolan, 1985[131]	RCT	3.5	Non-health: ratings of hotel services	Hotel guests (USA)	Self-completion survey	***Inclusion of space for free comment:*** Version 1: 10 items, no comments (200) Version 2: 5 items, open comment (200) Version 3: 5 items, structured comments (200) (Also manipulated: length of questionnaire) See also *Table 9*	Instrument response rates Mean ratings for common items Provision of additional comments	Instrument response rates: Version 1 – 12%; version 2 – 22%; version 3 – 24%; versions 2 + 3 combined – 23% RR = 0.49 (95% CI, 0.31 to 0.77) version 1 vs. version 3 RR = 0.52 (95% CI, 0.33 to 0.83) version 1 vs. version 2 RR = 0.94 (95% CI, 0.65 to 1.34) version 2 vs. version 3 RR = 0.51 (95% CI, 0.33 to 0.77) version 1 vs. versions 2 + 3 combined Mean ratings for common items: Version 1, 20.9; version 2, 23.1; version 3, 23.4 (Significance not reported and insufficient data to calculate CIs) Provision of additional comments: Version 2, 43%; version 3, 24% RR = 1.85 (95% CI, 0.99 to 3.42) version 2 vs. version 3

continued

TABLE 8 contd *Response categories*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Criteria for comparison	Main findings
Bishop *et al.*, 1986[92]	RCT	4	Non-health: views about social issues	Householders with telephones (USA)	Telephone survey	***Inclusion of DK response category:*** Form A: filter allowing respondents to indicate they had not thought about issue (397–408) Form B: no filter, but respondents allowed to give DK response (395–411) Form C: no filter, respondents pressed to select substantive response (397–404) 2 replications, with 3 fictitious topics/replication: different achieved sample sizes/ replication and topic (Also manipulated: ordering of fictitious and genuine issues; question wording (use of filter questions)) See also *Table 6*	Expression of opinion (% expressing an opinion)	% respondents offering an opinion on fictitious issues significantly higher when DK response probed (i.e. form C) ($p < 0.05$ for all comparisons) (Because of multiple comparisons, space constraints preclude presentation of RRs and 95% CIs)
Bishop, 1987[132]	RCT	4	Non-health: views about social issues	Householders with telephones (USA)	Telephone survey	***Inclusion of "middle" response category:*** Group sizes ranged across series of studies, but precise information not given Each study involved 2 groups; in 1 group, a middle alternative response category was explicitly offered by the interviewer; in the other, the middle alternative was not offered but was accepted if volunteered by the respondent	Distribution of responses (% respondents choosing "middle" category)	% respondents choosing middle alternative; for all 12 comparisons was significantly greater when it was explicitly offered ($p < 0.0001$ in all cases) (Because of multiple comparisons, space constraints preclude presentation of RRs and 95% CIs) Mentioning a middle alternative in question (but not offering it in the list of response categories) also increased % respondents choosing middle alternative ($p < 0.001$ for 2/3 items) (Because of multiple comparisons, space constraints preclude presentation of RRs and 95% CIs) Order in which "middle" alternative offered (second or last position) affected response; tendency for % opting for "middle" alternative to be greater when offered last in list of response categories, but direction of this ordering effect not invariable (Because of multiple comparisons, space constraints preclude presentation of RRs and 95% CIs)

continued

TABLE 8 contd *Response categories*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Criteria for comparison	Main findings
Wandzilak *et al.*, 1987[133]	RCT (cross-over design)	4	Non-health: attitudes to sport	Adolescent males (USA)	Self-completion survey	***Inclusion of "middle" response category:*** Order A: 4-item scale first (56) Order B: 5-item scale first (56) Both scales were forced-choice agreement/disagreement scales; 5-item scale had mid-point of "undecided"	Distribution of responses on 5-item scale	For 3/8 items, null hypothesis (that those who opted for "undecided" category on 5-point scale would be equally divided between "agree" and "disagree" categories on 4-point scale) was rejected ($p < 0.05$) (Because of multiple comparisons, space constraints preclude presentation of RRs and 95% CIs) Proportion choosing "undecided": significant difference ($p < 0.05$) between A–B and B–A order for 2/18 items (Because of multiple comparisons, space constraints preclude presentation of RRs and 95% CIs)
Ayidiya and McClendon, 1990[117]	RCT	4	Non-health: views on social issues	Householders with telephones (USA)	Postal survey	***Inclusion of "middle" response category; Ordering of response categories; Labelling of response categories:*** Version 1 (266) Version 2 (266) Questionnaire versions differed with respect to: order in which questions were presented; order of response categories; inclusion of "no opinion" filter questions; provision of middle alternative response categories; agree–disagree vs. forced-choice options (Also manipulated: question ordering) See also *Table 7*	Distribution of responses	Significant question order effects ($p < 0.05$) for 1/4 comparisons; another approached statistical significance ($p = 0.07$) Tendency towards primacy effect ($p = 0.08$) for response categories for 1/3 comparisons No evidence of recency effects (p = ns) for response categories for any of 3 comparisons % replying DK significantly higher ($p < 0.01$) with inclusion of "no opinion" filter for all 4 items considered % choosing "middle alternative" significantly greater ($p < 0.01$) when this option explicitly offered % agreeing with statement greater for "agree–disagree" format than for forced-choice format; difference significant for 2/3 items ($p = 0.03$ and $p < 0.01$ respectively) (Because of multiple comparisons, space constraints preclude presentation of RRs and 95% CIs)

Chapter 5

Questionnaire appearance

Introduction

Expert opinion, as well as common sense, tells us that attention to the appearance of a questionnaire, including its length and layout, is important. As Sudman and Bradburn[7] noted, in interview surveys a well-designed questionnaire can simplify the tasks of both interviewers and data processors. Through good design, the risk of errors in posing questions and coding responses can be reduced and potential variability between interviewers or coders can be minimised, thus reducing bias. In a self-completion questionnaire, its appearance is one of many factors influencing a recipient's decision to respond; response rates can be enhanced and the potential for non-response bias reduced by careful attention to questionnaire appearance and format. A "user-friendly" format can also reduce the risk of bias arising from respondent error (e.g. incorrectly skipping questions that should be answered, or ticking the wrong box).

The theory of social exchange[142–144] suggests that the actions of individuals are influenced by the rewards they expect to obtain from completing these actions and the costs of doing so. This theory underpins Dillman's[1] Total Design Method for postal and telephone surveys and was espoused by Brown and colleagues[140] in their model and analysis of responses to mail surveys. In this "task analysis" model the authors suggested that issues of questionnaire appearance can influence respondents' decisions at several stages in the decision-making process. The first stage is to arouse interest in the task of questionnaire completion; an attractive appearance can help here. The second stage is evaluation of the task; perceptions of the time and effort required to complete the questionnaire may be influenced by aspects of the questionnaire's appearance. The third stage is initiation and monitoring of the task of completion; here the actual burden of response becomes apparent and once again issues of length and format may come into play.

Sudman and Bradburn[7] proposed that the needs of three parties – the respondent, the interviewer and the data processor – should be taken into account in designing and formatting a questionnaire. They argued that the needs of the respondent should always be afforded the highest priority, followed by those of the interviewer; data processors are typically operating under less pressure than interviewers, so their needs should be given lowest priority. In contrast, de Vaus[50] argued that the relative weight given to the needs of each party should depend on the mode of questionnaire administration. In particular, he suggested that the primary concern in designing questionnaires for use in telephone surveys should be the convenience of the interviewers; they should be assisted to administer the questionnaire accurately and (if possible) to code responses as the interview proceeds. In general, the authors of the current review are inclined to agree with this latter viewpoint and to afford highest priority to the individual responsible for finding the way through the questionnaire and for recording the responses (i.e. respondents for postal or other self-completion questionnaires and interviewers for face-to-face or telephone interview surveys). No studies were identified that compared a version of a questionnaire explicitly designed to give priority to the needs of, for example, interviewers with a version designed to give priority to, for example, data processors. However, practitioners are well aware of the needs to reconcile these needs.

Issues of questionnaire appearance and layout that have been discussed in the literature are summarised in *Box 3*. Within the classic texts on survey methods and questionnaire design, however, issues of questionnaire formatting have received far less attention than those of question wording and sequencing. Moreover, the limited recommendations made in these books are largely based on the opinions and experiences of the authors, rather than on theories of perception, cognition and response behaviour, or on the systematic study of aspects of layout. In the words of Jenkins and Dillman,[145] there have been "few systematic efforts … to derive principles for designing self-administered questionnaires from relevant psychological or sociological theories". In this review, too, relatively few comparative studies of the impact of questionnaire appearance and design on response rates and response bias were identified; the authors have therefore supplemented the results from such comparative studies with findings from earlier literature reviews and with expert opinion.

BOX 3 Issues of questionnaire appearance and layout

- Length of questionnaire
  - Number of questions
  - Number of pages
- Pagination
  - Use of booklet format
  - Page size
  - Double- versus single-sided printing
  - Placement of questions within pages
  - Use of "white space"
- Paper colour and quality
- Print details
  - Font size
  - Typeface
  - Print colour
- Cover design
- Question and response category format
  - Identifying questions
  - Vertical versus horizontal response formats
  - Placement of codes
  - "Tick box" versus "circle number" response formats
  - Indication of skip and branch patterns
- Instructions
  - Types of instructions
  - Placement of instructions

Length of questionnaire

The length of a questionnaire may be conceptualised in many ways:

- the number of (numbered) questions
- the average number of data items requiring response (which may be greater than the apparent number of questions because of subsidiary questions, or may be fewer because of filtering and skipping)
- the number and size of the pages
- how long an interviewer says an interview will take (or how long a self-completion questionnaire is stated to take in a covering letter)
- some function of the preceding four criteria, representing "perceived respondent burden", as postulated by Brown and colleagues[140]
- how long an interview (or filling in a self-completion questionnaire) actually takes
- the cognitive load imposed on the respondent (e.g. difficulty of the concepts, how complicated the layout is, whether difficult acts of recall or arithmetic manipulations are required)
- some function of the preceding two criteria, representing "actual response burden".

In practice, length appears to have been conceptualised, operationalised and measured largely as one of the first four of the above criteria.

It is generally held that response rates are inversely related to the length of the questionnaire (however defined). A "survey of surveys"[146] provided some weak evidence that response rates are inversely related to the length of the interview.

The assertion of an inverse relationship is underpinned by the general theory of respondent motivation. For example, Cannell and Kahn[147] suggested that, in interview surveys, motivation may decline once the interview is extended beyond some optimal length. However, Bradburn[148] argued that when the survey is perceived to be important or interesting to respondents, the survey instrument can be quite long without having a detrimental effect on the rate or quality of response.

A point that is also of relevance, but receives less attention in the literature, is whether response error and response bias are a function of questionnaire length, although Houston and Ford[149] have suggested that the scope of research on survey methods needs to encompass these aspects of "response quality". Sudman and Bradburn[7] suggested that response bias in favour of individuals with strong negative or positive opinions on the topic are likely to occur in respect of long questionnaires. Furthermore, in long questionnaires or interviews, fatigue and boredom may lead to individuals becoming careless or adopting response strategies that reduce the burden of answering, especially with respect to questions in the latter part of the questionnaire.

Identified studies

Eleven randomised controlled trials or non-random concurrent controlled studies[47,80,131,150–157] in which questionnaire length had been manipulated were identified. In all of these, self-completion questionnaires were used (*Table 9*; see p. 94). The specific aspects investigated were:

- the impact on questionnaire response rates[47,131,150–157]
- patterns of response to individual questions[80,155,156]
- item non-response rates[151]
- speed of response.[151]

All studies except one were conducted as randomised controlled trials; in the remaining study,[156] a non-random concurrent controlled design with statistically comparable controls was employed.

Only two studies[47,150] were on health-related topics; this may affect the generalisability of findings because surveys on health-related topics typically achieve better response rates than those on more general issues. None reported that sample sizes were based on a power calculation. Quality scores for several studies were also affected by the fact that factors other than those under immediate scrutiny were not held constant. However, the authors of the original articles generally argued that there were practical reasons for the lack of consistency of treatment (e.g. the difficulty of devising a long questionnaire using only factual questions[47]).

From *Table 9* (see p. 94) it is clear that the definition of "long" and "short" questionnaires varied widely between studies. Length was generally defined and operationalised in terms of the number of questions and pages. However, the "long" version in some studies was in fact shorter than the "short" version in others. Furthermore, in many of the studies, other variables that were expected to influence response rates and patterns (e.g. mode of contact, types of questions included) were also manipulated. In addition, settings and topics were heterogeneous. For these reasons no attempt was made to combine results across studies. Further-more, apparently contradictory findings across studies may be related to this lack of homogeneity.

Response rates

Evidence from primary studies

The findings from the identified studies with respect to the impact of questionnaire length are equivocal. In the two studies with a health focus the researchers found that length of questionnaire had no effect on response rates. Cartwright[47] compared three lengths of questionnaire eliciting new mothers' experiences of pregnancy and labour and found no significant differences in overall response. Jacoby[150] reported on a comparison of "long" and "short" versions of a questionnaire in the context of a study of patients' views and experiences of general practice. No significant differences in response rates between the two versions were found. In both of these studies other variables hypothesised to influence response rates (e.g. questionnaire content, sponsorship) were also manipulated; the finding of no relationship between response rate and questionnaire length remained when these other factors were held constant.

In three studies, on a range of non-health topics,[151,154,155] researchers found significantly lower response rates for longer questionnaires. However, Hansen and Robinson's[151] data suggest that this effect may be modified by the intensity of prenotification; the CI for the RR of long versus short questionnaires includes unity when controlling for those respondents who were contacted prior to receiving the questionnaire. In the study by Powers and Alderman,[154] the finding of lower response rates for longer questionnaires held true on controlling for offer of feedback of survey findings.

Adams and Gale[153] found that response rates for a medium length questionnaire were significantly better than those for either a short or a long questionnaire; response rates to the longest version were significantly lower than those to the medium and short versions.

The results from two studies on the topic of hotel guests' satisfaction with services (a population and topic where response rates are typically low) provided mixed evidence on whether response rates to longer questionnaires were lower or higher. Trice and Dolan[131] reported lower response rates for a ten-item questionnaire compared with a five-item questionnaire (12% versus 23%) in the first of their studies, which was a significant difference. However, length of questionnaire and the provision of space for additional comments may have been confounded in this study; the ten-item questionnaire did not provide any opportunity to make additional comments, while the two five-item versions did (half with provision for open comments, half with provision for structured comments). In contrast, in a second survey reported in the same article, Trice and Dolan[131] found a trend towards increased response rates with increasing numbers of questions, although the differences did not reach statistical significance. In a later study, Trice[157] noted that response rates to short questionnaires with no space for comments were lower than those to longer questionnaires and to short questionnaires with space for comments.

In all of the studies described above, the "long" and "short" versions varied with respect to both the number of pages and the number of questions; it is therefore difficult to ascertain whether the respondents were reacting to the number of items (the actual burden of response) or the number of pages (the perceived burden). Layne and Thompson[152] compared two versions of the same 30-item questionnaire; in the "long" version, the questions were spread over three pages while in the "short" version, they were concentrated on one page. Contrary to their expectations that the "long" version would result in poorer response rates, they found no significant difference between the two versions. They concluded that the longer version

may have been more "aesthetically appealing" because in the short version the questions were crammed together.

Evidence from earlier reviews
In addition to the primary studies described above, the search identified six earlier literature reviews that included a consideration of the impact of questionnaire length on response rates.

Linsky,[158] in a review similar to the current report, identified seven studies (dating from the 1940s to 1970) in which length of questionnaire had been manipulated. In four of these, no significant differences in response rates between long and short versions were found. For one study, the response rate for a short (postcard format) questionnaire was 28% higher than that for a two-page questionnaire, while, for the remaining two studies, substantially higher response rates were obtained for the longer version. Linsky[158] recognised, however, that most of the studies in which length of questionnaire was manipulated failed to control for potentially important confounding factors.

A similar review by Kanuk and Berenson[159] also identified seven studies in which the length of the questionnaire had been manipulated; these included five of the seven identified by Linsky.[158] In four of the seven studies, no significant impact of questionnaire length on response rates was detected, while in a fifth the significance of findings was not reported but the original researcher concluded that the "necessity of using short questionnaires with mail panels was more folklore than fact". In the sixth study, a postcard with a single question was compared with a three-page questionnaire; response rates to the shorter version were 22% higher. In the final study, adding one or two pages to questionnaires that were already three to six pages in length had no impact on response rates, but including additional "interesting" questions to a largely "uninteresting" questionnaire increased response rates. As with Linsky,[158] Kanuk and Berenson[159] recognised the possibility of confounding of length of questionnaire with other factors.

In their review and meta-analysis, Heberlein and Baumgartner[160] found no significant zero-order association between rate of response and length of questionnaire (whether measured in terms of the number of questions, number of pages or time required to complete the questionnaire). However, on controlling for saliency (a "salient" topic being defined as "one which dealt with important behaviour or interests that were also current") and the number of contacts made with the respondent, an inverse relationship between number of items and response rate was demonstrated, with each additional question reducing the response rate by 0.05%. Goyder[161] sought to replicate the model developed by Heberlein and Baumgartner,[160] but included additional material. In his multivariate predictive model, questionnaire length (measured in number of pages) was inversely related to response rates.

In their review of factors influencing response rates to surveys on leisure and natural resources topics, Brown and colleagues[140] found that the number of pages in the questionnaire was one of the five variables to enter a multiple regression model; response rates were inversely related to the number of pages. However, font size was another explanatory variable, with a higher response rate related to a larger type-face. They concluded that a slight net advantage (of approximately 4%) could be achieved by using a 16-page questionnaire with 4/32" type rather than an eight-page questionnaire with 2/32" type.

Finally, Yu and Cooper[162] found a negative but extremely weak ($r = -0.06$) association between number of questions and response rate.

Item non-response rates
Hansen and Robinson[151] found no significant differences in mean percentages of unanswered items for long versus short questionnaires; this held true for all respondents together and on controlling for intensity of prenotification.

Response patterns
Roszkowski and Bean[155] did not find any evidence of "response bias" (gauged in terms of the distribution of responses to a scale measuring satisfaction with an educational course) due to questionnaire length.

Herzog and Bachman[156] found that "straight line" responding (i.e. endorsing the same response category) was more likely in the latter part of a long questionnaire than in a short questionnaire; the exception was in relation to questions of personal relevance to the respondents. They suggested that this possibly indicated fatigue or carelessness in responding to these later questions and may therefore have represented a source of response bias. However, their results showed that "straight line" responding tended to occur within an "item set" (i.e. a series of related questions, generally with the same response categories) rather than across all questions, suggesting that the bias was largely restricted to related items.

Helgeson and Ursic[80] interviewed respondents after their completion of electronic and pencil-and-paper questionnaires and probed the cognitive processes they had used (i.e. how they had interpreted the questions, and arrived at and recorded their answers). They postulated that differences in conceptualising and framing answers may be a source of response bias. For the electronic version there were no differences with respect to questionnaire length but, for the pencil-and-paper questionnaire, reported cognitive processes varied between the short and long versions. This finding suggests that the potential for response errors may vary according to mode of administration. However, the interaction effect was significant only at the 0.10 level. Furthermore, the cognitive processes affected accounted for only 19% of all coded thought processes and related mainly to statements of brand names, an issue of little relevance in surveys of patients and health professionals.

Speed of response

Although it may be expected that respondents would take longer to complete and return a longer questionnaire, Hansen and Robinson[151] found that length of questionnaire had no significant impact on the speed of response; the main determinant was whether prenotification of the survey had occurred.

Conclusions from identified studies

Roszkowski and Bean[155] concluded that increasing the length of a questionnaire is likely to have an adverse effect on response rates when the difference in length between the short and long versions is sufficient to place a significantly greater burden on respondents and when questionnaire salience is low. Herzog and Bachman[156] concluded that, in long questionnaires, item sets present an opportunity for respondents to reduce the burden of response by adopting a uniform response strategy. In general, however, the evidence on the relationship between questionnaire length and response rates was equivocal. Moreover, saliency (relevance and interest) appeared to be a moderating factor, as previously noted by Kanuk and Berenson.[159] Surveys of patients and health professionals may be perceived by the target respondents as having relatively high saliency, an assertion supported by the negative findings with respect to the impact of questionnaire length on response rates by Cartwright[47] and Jacoby;[150] a relatively long questionnaire on a health-related topic may therefore be acceptable. Nonetheless, common sense, supported by findings from previous literature reviews by Heberlein and Baumgartner[160] and by Goyder,[161] suggests that, even for surveys on health topics, there may be an upper limit of length beyond which response rates and response quality are likely to decline. This limit may of course depend on many other factors, such as the population surveyed, the survey topic, and other aspects of questionnaire appearance. For this reason (and bearing in mind resource costs and ethical considerations), superfluous questions should be avoided. The use to which each piece of information gathered will be put should be clear from the outset and the aim should be to collect information that is "necessary to know" rather than "nice to know".

Pagination

Few studies on this aspect of questionnaire appearance were identified. Suhre[163] considered questionnaire size, while other researchers[131,152,157] examined the amount of "white space" provided. All were randomised controlled trials but none was on a health-related topic. For the most part, textbook recommendations on pagination are based on the expert opinion and experience of the authors concerned.

Use of booklet format

Printing the questionnaire as a booklet (i.e. on large sheets of paper folded and, if necessary, stapled through the spine) has been recommended explicitly by Dillman[1] and by Sudman and Bradburn,[7] who suggested that this format provides greater ease in reading and turning pages, reduces the risk of losing pages, and facilitates the use of a double-page format for questions about multiple events or persons. A booklet format was also favoured by Bourque and Fielder,[40] who argued that it presented a "more professional" appearance. No studies that compared a booklet and another format were identified.

Size of page

As part of his Total Design Method, Dillman[1] advocated the use of 12.25 × 8.25 inch paper, folded to produce a booklet of 6.125 × 8.25 inches. The exactness of these dimensions was to ensure that the folded questionnaires fitted readily into US monarch size (7.25 × 3.875 inches) envelopes and regular size (6.25 × 2.75 inches) business reply-paid envelopes, and could be mailed (together with a covering letter and return envelope) for minimum first class postage; he emphatically rejected use of the more widely available 8.5 × 14 inches legal-size paper because the slightly larger format booklet would be likely to tip the postal costs into a higher bracket. Bourque and Fielder,[40] however, suggested the use of 8.5 × 17 inches paper (folded), to facilitate the use of larger print.

Identified studies

Only one study[163] meeting the quality criteria and comparing the effect of different sizes of questionnaire was identified (*Table 10*; see p. 98). In this, the dependent variable was response rate. There were no significant differences in response rates between A5 and A4 format questionnaires (controlling for modes of advance notice and of follow-up and presence/type of incentive).

Double- or single-sided printing

Scant attention has been given to this topic. De Vaus[50] recommended that questionnaires should be printed on one side of the page only, arguing that respondents may miss questions printed on the "backs of pages" and suggesting that the blank pages may be useful for respondents to provide additional information. Implicit within the recommendation of a booklet format,[1,7,40] however, is the notion that questions are printed on both sides of the page. No studies that explicitly addressed double-sided versus single-sided print formats were located.

Placement of questions within pages

The issue of question sequencing is discussed in chapter 4. The placement of questions within pages is also of importance. No studies were identified in which question placement was manipulated or examined. However, most basic texts on questionnaire design and the conduct of surveys provide guidance on this topic. Dillman[1] stated that "the pages must be constructed in a way that keeps respondents from skipping individual items or whole questions". Having to turn a page in the middle of a question is confusing and likely to give rise to response errors. In particular, having the question on one page and the associated response categories on another is to be avoided if at all possible. As a general principle, Dillman[1] recommended that text should be organised so that the question, associated instructions and response categories all appear on a single page. Bourque and Fielder[40] suggested that splitting response categories over two pages may introduce a subtle form of loading because respondents may not read or may give less consideration to the items on the second page. Dillman[1] and Bourque and Fielder[40] all recommended that, if it is impossible to fit the entire question and response categories on one page (as may be the case in those involving ratings of a series of items), every effort should be made to place the question stem on a left-hand page, with the list of items or response categories continued on the facing page; the placement of response categories relative to question stems is discussed in greater detail in chapter 4.

Dillman[1] also cautioned against asking the respondent to do two things at a time, for example, to rate the strength of agreement with a series of statements and to rank these statements in order of importance. Instead, he recommended using two separate questions in such situations. This approach requires either that the items in the series are repeated in each question or a cross-reference is made between the two questions, for example, by using numerical or alphabetical labels to identify the items in the series. If a cross-referencing approach is chosen, the two questions should be placed on the same or facing pages. Bourque and Fielder[40] highlighted the desirability of placing a question that is logically dependent on a previous question on the same page as that predecessor. Sudman and Bradburn[7] cautioned against having a long question with a number of subparts followed by a short question at the foot of the page; it was their experience that a short question in this position was often omitted in error.

Dillman[1] suggested some strategies for preventing questions being split over two pages. One is to reorder questions within the questionnaire, if this can be done without breaching the principles of question sequencing as set out in chapter 4. Another is to adjust the spacing of questions, preferably by manipulating the amount of "white space" between them or by adjusting the margins. Sudman and Bradburn[7] advocated the use of parallel columns and facing pages when asking identical questions about multiple events or persons (e.g. about the health of all household members). They suggested that, if questions about a particular person or event covered more than one page, die-cut (shortened) pages could be used so that the identifying information was always visible.

Use of "white space"

Sudman and Bradburn[7] recommended against crowding questions in the hope that this strategy would make the questionnaire look shorter! They argued that an apparently longer questionnaire with a less cramped layout and more "white space" looks easier to complete, generally resulting in higher response rates and less response errors. In particular, they highlighted the need to leave sufficient space for responses to open-ended questions, stating that "the answer will not be longer than the space provided". Interviewers and respondents to self-completion questionnaires both take the amount of space as an implicit indication of the level of detail required in a response. These authors also cautioned against the use of lines for open-ended questions, arguing that they make the questionnaire look more crowded. Their exception

to this principle was in respect of questions for which only a short answer (i.e. one or a few words, or a number) was required; in these circumstances they recommended that a line should be used.

Identified studies

Three studies were identified[131,152,157] in which aspects of white space were manipulated. All three were randomised trials on non-health topics and all examined the impact on response rates. Two also examined the impact of providing space for open comments on the volume and quality of individual responses. Because the provision of more or less white space inevitably affects the length of the questionnaire, these studies are also discussed under "length of questionnaire" above and the findings are summarised in *Table 9* (see p. 94).

Response rates

Findings on this issue are equivocal. Layne and Thompson[152] found no difference in response rates to one-page and three-page questionnaires when the number of questions was held constant; this finding was contrary to received wisdom that longer questionnaires achieve poorer response rates, so they concluded that the less cluttered appearance of the longer questionnaire (i.e. having more white space) was more aesthetically appealing. Trice and Dolan[131] found that providing space for additional comments increased response rates. In contrast, in his later study, Trice[157] reported that providing an explicit space on the questionnaire for open comments had a non-significant impact on overall response rates when controlling for number of items. However, the response rates to a short questionnaire with space for open comments were broadly similar to those for a longer questionnaire (whether with or without space for comments) and were significantly higher than those for a short questionnaire with no such space.

Volume and quality of response

Trice and Dolan[131] also showed that more comments were provided in an unstructured condition (i.e. when the respondents were free to choose the aspects of service upon which to comment compared with providing structured headings for comments), but the observed difference just failed to reach statistical significance.

Paper colour and quality

Only one study was located that dealt specifically with this feature of questionnaire appearance, a randomised controlled trial on a non-health topic.[164] Once again, textbook recommendations are based almost entirely on the expert opinion and experience of the authors concerned.

Dillman[1] favoured white or off-white paper, although his reasons were not stated. Sudman and Bradburn[7] suggested, however, that the use of coloured covers or coloured sections "may be helpful to interviewers when multiple forms are used or for complex skipping patterns". Nevertheless, they recommended against the use of dark-coloured papers because these are more difficult to read. Likewise, Bourque and Fielder[40] advocated that there should be a good contrast between print and paper. They cautioned against the use of neon colours or those that reduce the contrast for colour-blind people and suggested that, "when in doubt, use black print on a white background". Jenkins and Dillman,[145] while recognising the paucity of evidence on the impact of colour on response rates or response quality, suggested that sophisticated tone-on-tone colour schemes (e.g. deep blue print on a light blue background, with white boxes or spaces to highlight where responses should be made) could be used to define "the desired navigational path".

In the UK a number of organisations have produced guidelines on enhancing the readability of documents, although not specifically with reference to questionnaires. The Basic Skills Agency (whose remit is assisting those with literacy problems) recommends the use of paper that is sufficiently thick to avoid marked "shadowing" of the text from the previous page. The Agency also comments that a dark background is generally more difficult to read from and that certain colours, notably blue and purple, are worse than others. The Royal National Institute for the Blind states that the contrast between the type and the paper is important in determining legibility; black type on a white or yellow background gives the best contrast; reversed-out type (e.g. white on a black or other dark background) is also acceptable.

Identified studies

Only one study[164] was identified that met the quality criteria and in which the colour of the paper was manipulated; the dependent variable was response rate (*Table 11*; see p. 98). In this study the authors compared identical questionnaires printed on blue and white paper and found no significant difference in overall response rates. A single reminder, including a copy of the appropriately coloured questionnaire was sent approximately 4 weeks after the initial mailing; response rates to the blue and white questionnaires were not significantly different before or after this reminder.

Print details

This is yet another topic on which empirical evidence is lacking and reliance has to be placed on the expert opinion of survey researchers and others with a professional interest in typography.

Font size and typeface

Dillman[1] recommended the use of 12-point Elite typeface, subsequently photographically reduced to 79% of the original size; he gave no explicit justification, but it should be noted that he was writing at a time when typewritten rather than word-processed production of questionnaires was the norm. Sudman and Bradburn[7] were less prescriptive, simply suggesting that the typeface should be large enough and clear enough to avoid strain in reading. They postulated that the exception to this rule should be instructions intended solely for data processors; these can and indeed should be put in a smaller type, to avoid distracting the interviewer or the respondent. Bourque and Fielder[40] recommended a 10-point font size and a font with equal character spacing, such as Courier, to avoid problems of alignment.

Bourque and Fielder[40] recommended using a combination of bold, underlining and upper case to provide emphasis in the text of a question and to distinguish instructions from questions. They advocated against the use of an italic font because this may be difficult to read.

The Basic Skills Agency suggests that typefaces need to be clear and distinct (for example, typefaces in which "rn" can be mistaken easily for "m" should be avoided). They state that the font size should be related to the nature and purpose of the text and that the leading (space between the lines) should in turn be related to the font size; for example, with a font size of 12 points, leading of 2 points is desirable. They also suggest that the overuse of upper case letters, for example to convey emphasis, is counter-productive and that the use of bold type or boxing is more appropriate in these circumstances. The Royal National Institute for the Blind recommends a font size of 12 points for documents intended for general readers, and a minimum of 14 points if readers are likely to have a visual impairment. Both these organisations express a slight preference for san serif fonts (such as Arial).

Line spacing

Bourque and Fielder[40] recommended using at least double spacing between one question and the next, and between a question and the first related response category; within a set of response categories, they proposed the use of 1.5 line spacing.

Cover design

No studies were identified on this aspect of questionnaire appearance. However, Dillman[1] stressed the need for particular attention to the front and back covers of questionnaires and recommended that the front cover should contain: the title of the survey (which should convey its purpose in an "interesting but neutral" manner), the identity of the organisation carrying it out, some form of graphic illustration, and brief instructions. Sudman and Bradburn[7] largely concurred with this, especially in respect of postal questionnaires, although they suggested that the illustration could be omitted in surveys of professional groups and for short (two-page) question-naires. Both Dillman, and Sudman and Bradburn, advocated the use of a "neutral" illustration.

Sudman and Bradburn[7] recommended that the outside back cover of the questionnaire should be left blank, with the invitation that respondents can use this space for any additional comments. Dillman[1] also suggested the provision of an invitation to make additional comments and a "thank you" to respondents.

Question and response category format

Jenkins and Dillman[145] sought to develop a theory of self-administered questionnaire design. They emphasised the need for an understanding of "graphic non-verbal language", in other words, the spatial arrangement of information and other visual phenomena such as colour and brightness. Drawing on theories of cognition, perception and pattern recognition/processing, they argued the need for consistency in the presentation of visual information and derived five principles of design for self-administered questionnaires.

1. "Use the visual elements of brightness, color, shape and location in a consistent manner to define the desired navigational path for respondents to follow when answering the questionnaire."…
2. "When established format conventions are changed in the midst of a questionnaire, prominent visual guides should be used to redirect respondents."…

3. "Place directions where they are to be used and where they can be seen."…
4. "Present information in a manner that does not require respondents to connect information from separate locations in order to comprehend it."…
5. "Ask people to answer only one question at a time."[145]

As Jenkins and Dillman[145] recognised, not only is there a lack of theoretical underpinnings to issues of questionnaire format and design but there have also been few studies of any aspect of question/ response category format. Only one such study that met the criteria[47] was identified (described under "tick box" or "circle number" response formats below). Otherwise, authors have drawn on extensive experience of survey design and administration in making their recommendations, which are set out in greater detail below.

Methods of identifying questions

Dillman[1] recommended the use of case to distinguish questions from response categories, advocating the use of lower case letters (which are more readable) for the questions and upper case letters for the response categories. Sudman and Bradburn[7] advocated numbering each question to minimise the risk of questions being skipped, to facilitate cross-referencing and the use of skip instructions, and for ease of reference in data processing. They recommended the use of Arabic numerals to identify main questions, with subparts being denoted by letters and, if necessary (i.e. if there are further subdivisions), by numerals placed in parentheses. They also proposed that subparts of questions should be indented.

Vertical versus horizontal response formats

Dillman,[1] and Sudman and Bradburn[7] all recommended the use of a vertical answer format for individual questions, suggesting that: this is somewhat easier for interviewers, data processors and especially respondents to self-completion surveys; it gives a less cluttered appearance; and it provides white space for interviewers or respondents to include additional comments adjacent to the appropriate question. Dillman[1] also argued that the vertical flow pattern adds to a "respondent's feeling of accomplishment". Bourque and Fielder[40] echoed this view, postulating that it differentiates the response categories from the question and from each other. Jenkins and Dillman[145] offered a theoretical basis for a vertical presentation of response categories in multiple choice closed-ended questions. They argued that the Gestalt Grouping Laws and the literature on graphic language indicate that a vertical presentation gives the correct impression that the categories are distinct entities.

An exception to the rule of vertical format has been suggested in the case of a set of questions that all use the same response categories (e.g. a Likert scale). In this situation, all of the above authors recommended a horizontal format, primarily on the grounds of conservation of space. Jenkins and Dillman,[145] drawing on theories of pattern recog-nition and processing and on the experimental work of Gaskell and colleagues[166] argued that presenting scaling categories in a vertical format may give the erroneous impression that each category is independent of the others. A horizontal presentation, in contrast, enhances the perception of an underlying continuum to the response scale.

Dillman[1] also recommended numbering the items in the scale, indenting the second and subsequent lines of the statements to be rated, aligning the response categories with the last line of the corresponding statement, using leading dots from the end of each statement to the beginning of the corresponding response categories and using a "hat" (bracketing) over the columns of response categories.

Placement of headings

Dillman,[1] and Sudman and Bradburn[7] advised against placing headings in a sideways format (i.e. at 90° to the remainder of the text); instead, they suggested that headings could be made to fit in a horizontal format simply by using more space. (The labelling of response categories is discussed in greater detail in chapter 4.)

Placement of codes

Dillman[1] favoured the use of numbers (to be circled) over dashes or boxes (to be ticked), arguing that this provided a convenient form of precoding and thus facilitated the task of the data processor. He stated that the numbers should always be placed to the left of the response category to maintain consistency of spacing when some response categories are longer than others and to allow space to the right of the response category for the respondent to supply any additional information required (e.g. in response to a prompt of "other, please specify"). Although not indicating any preference for placing the codes to the right or the left of the associated descriptors, Oppenheim[13] stressed the importance of consistency in their placement. Bourque and Fielder[40] favoured placing the response codes to the right of the word or phrase to which they refer because English is read from left to right. They argued that this also facilitates the task of data entry and recommended right-aligning codes and using

leading dots to aid respondents in linking the correct response to the corresponding code. (The labelling of response categories is discussed in greater detail in chapter 4.)

"Tick box" versus "circle number" response formats

As noted above, Dillman[1] favoured a "circle the number" format in the interest of data processing and because, at the time of writing, producing boxes was more difficult. As with the placement of codes, Oppenheim[13] stressed the need for consistency in the method of answering multiple choice questions; in other words, a mixture of circling and ticking should be avoided.

Identified studies

Cartwright,[47] in a study of new mothers' attitudes to the management of pregnancy and labour, compared "tick the box" and "circle the number" formats (*Table 12*; see p. 99). The former had been used in previous waves of the survey and had necessitated time-consuming and expensive post-coding. There was no significant difference in response rates or response quality, as measured by inadequate responses, between the two formats.

Indication of "skip" patterns

Interviewers and respondents can be guided to the appropriate question either by verbal instructions or by arrows pointing to the question. Sudman and Bradburn[7] reported that, in their experience, verbal instructions are adequate. They also recommended that skip instructions should be placed immediately after the answer giving rise to the skip and that skips should be positively rather than negatively worded (i.e. skip if a particular answer is given rather than skip if an answer is not given). Dillman,[1] however, believed that relying on words alone to describe routing through the questionnaire was unacceptable. Instead, he recommended the use of arrows to direct respondents from the screening or filter question to the next applicable question, the indentation of conditional questions, and the use of boxes to direct respondents past questions that are inapplicable to them. When all respondents are to be asked the same number of questions but the actual questions to be asked depend on their responses to a previous question, Dillman[1] suggested that a vertically split page may be used.

Instructions

Type of instructions

Questionnaires and interview schedules may need to contain instructions and information for three different parties: the interviewer, the respondent and the data processor. Instructions for the interviewer are likely to include details of which questions are to be asked of which respondent and the "script" for interacting with the respondent, including the types of probe to be used in eliciting information. Instructions for the respondent can be conveniently subdivided into "general", "transitional" and "question answering".[40] "General" refers to introductory remarks about the purpose of the survey, the type of questions that are to be asked, why the information is required, and what should be done with the completed questionnaire. Dillman[1] suggested that "transitional" statements may be used in three situations: where there is a change in topic or line of inquiry; at the top of pages; and to break up the monotony of a long series of questions. "Question answering" instructions, as the name implies, provide guidance on how questions are to be answered (e.g. whether a number should be circled or a box ticked) and on routing (e.g. branch and skip instructions). Information for the data processor may include details of the fields or columns in which data should be entered into the computer program.

Placing of instructions

It is usually recommended that general information on what the survey is about and how the questions are to be answered (e.g. "most questions should be answered by circling one number") should be placed at the beginning of the questionnaire, while instructions pertaining to individual questions should be placed as close as possible to the relevant question.[7] Dillman[1] pointed out that it may not be necessary to repeat instructions for every question if the same mode of response is required throughout. However, he advocated repeating instructions where mixed modes of response were required (e.g. if some questions required more than one answer to be circled, while others required a single response to be endorsed). Sudman and Bradburn[7] suggested that, if the instruction is to do with who should answer the question, or how it should be asked (e.g. if the interviewer is to offer prompt cards to the respondent), it should precede the question; if it is to do with how answers are to be recorded or how the interviewer is to probe for information, it should follow the question. Instructions on what to do with the completed questionnaire should be placed at the end of the questionnaire. It is generally recommended[7,40] that the questionnaire should also end with a "thank you".

As an alternative to placing general introductory information (e.g. the purpose of the survey) in the questionnaire itself, such details may be given in a

covering letter or flier. However, there is a risk that respondents may ignore or overlook detailed instructions provided in a letter; for that reason, instructions specific to completion of the questionnaire should appear on the questionnaire itself.

Format of instructions

Dillman[1] advocated the use of parentheses and lower case letters to distinguish instructions in both questions and response categories. Similarly, Sudman and Bradburn[7] recommended the use of a distinctive type, such as upper case or italic, to distinguish instructions and probes (to be used by the interviewer) from the questions. With respect to interviewer-administered questionnaires, Fowler[86] highlighted the need to alert interviewers when the exact form of words requires a decision to be made (e.g. the choice of "husband" or "wife" as appropriate in a question about the respondent's spouse); parentheses or a different typeface can be used to highlight the need for such tailoring.

Conclusions

Few relevant studies are health related; the generalisability of findings to this field may be limited. For many of the topics investigated, it was possible to identify only one or two relevant studies that met the quality criteria. This means that caution should be exercised in interpreting and extrapolating findings. Nonetheless, in the absence of empirical evidence, a good deal of expert opinion makes sound sense and is supported by theories of cognition, perception and pattern recognition.[145] However, some of the recommendations made in the key texts are based on "old technology" (in particular, typewritten documents) and fail to take into account the facilities afforded by modern word-processing and desktop-publishing software, by current reprographics facilities, or by the use of scannable (OMR and OCR) questionnaires (see chapter 3).

Length of questionnaire

- Findings with respect to the impact of questionnaire length on response rates are equivocal.
- A saliency by questionnaire length interaction has been demonstrated in previous reviews; questionnaires on highly salient (relevant or interesting) topics (as will be the case in many surveys of patients or health professionals) can probably be longer than questionnaires on more general topics or those for a general population.
- There is the potential for response bias, due to fatigue or carelessness, in the latter part of long questionnaires, particularly with respect to answers to "item sets".

Pagination

- In terms of response rates, the superiority of a booklet format questionnaire, or of an A4 (as opposed to A5 or other size) format has not been demonstrated.
- Findings with respect to the provision of "white space" are equivocal.

Paper colour

- Questionnaire colour has not been shown to have a significant impact on response rates.

Question and response category formats

- Differences in response rates or response quality between a "tick the box" format and a "circle the number" format have not been shown to be significant.

Other aspects of questionnaire appearance

No evidence was identified from comparative studies on the following aspects of questionnaire appearance: double- versus single-sided printing; placement of questions within pages; print details; cover design; methods of identifying questions; vertical versus horizontal response formats; placement of headings and codes for response categories; identification of skip patterns; and nature, placing or format of instructions.

Recommendations for practice

Although no literature was identified on the topic, computer-scannable questionnaires (OMR and OCR) are likely to assume greater importance in the future. Design principles for scannable questionnaires should be guided by the hardware and software to be used.

Recommendations with an evidence base from one or more high-grade primary comparative studies

Length of questionnaire

- Avoid excessively long questionnaires, especially if the topic is likely to be of low saliency to the respondents. (Recommendation based on evidence from primary studies and previous reviews.)
- Avoid crowding questions or reducing "white space" in a desire to reduce apparent length. (Recommendation based on limited evidence from primary studies and on expert opinion.)

Response formats

- Use a "circle the number" format rather than a

"tick the box" format in self-completion questionnaires. (Recommendation based on limited evidence from one primary study and on expert opinion.)

Recommendations based on theories of perception and cognition and/or on expert opinion

In surveys of patients, it is likely that a significant proportion of the target sample will have some degree of visual impairment. The needs of such individuals should be taken into account in designing self-completion questionnaires.

Pagination

- Use a booklet format with double-sided printing.
- Use standard-sized paper (A4 folded to A5 booklet or A3 folded to A4 booklet, as dictated by length of questionnaire).

Placement of questions within pages

- Avoid splitting a question, its associated response categories and instructions for answering over two pages.
- In questions where the list of response categories is too long to fit on a single page, continue it on a facing page if possible; otherwise repeat the question on the subsequent page.
- Do not ask respondents to do two things at once (e.g. rating and ranking) in responding to one question.
- When one question is logically dependent upon another, make every effort to place both on the same page.
- Avoid placing a short question at the foot of a page, especially if preceded by a long question with a number of subparts.

Use of "white space"

- Leave sufficient space for responses to open-ended questions.
- Do not use lines for responses to open-ended questions, unless only a short response (i.e. a number or a few words) is required.

Paper colour

- Paper colour has not been shown to have a significant impact on response rates (evidence base), so choose white paper or a light tint to enhance legibility.
- Consider the use of coloured covers to distinguish questionnaires.

Print details

- Use a font size of at least 10 points; a larger font size (up to 14 or 16 points, depending on typeface) is desirable if it is anticipated that respondents may have some visual impairment (e.g. in surveys of older people).
- Use a distinct typeface and avoid excessive use of italics and upper case characters, especially in self-completion questionnaires.

Cover design

- The front cover of the questionnaire should contain the title of the survey (not "Questionnaire on X topic ..."), the identity of the organisation carrying it out and, for self-completion surveys, a neutral graphic illustration.
- The back cover should provide some blank space for respondents' open comments, and should specify the address of the organisation conducting the survey (if not on the front cover) and say "thank you" to the respondent.

Question and response category format

- Use elements of brightness, colour, shape and location to "steer" the respondent through the questionnaire.
- Maintain a consistent format throughout the questionnaire.
- Use a vertical response format (*Figure 1*) for closed questions, except for rating scales.
- Use a horizontal response format (*Figure 2*) for item sets involving the same response categories throughout, and in rating scales.
- Consider natural reading style (i.e. left to right, and horizontally orientated) in placing headings and codes for responses.
- Use graphical means (e.g. arrows and boxes) to indicate skip patterns.
- Place instructions and directions at the point where they are required; if a series of questions involves turning a page, it may be necessary to repeat instructions on the new page.

Recommendations for future research

As already noted, issues of questionnaire format and appearance have been under-researched to date. The time is therefore ripe for studies on the impact of questionnaire design. The authors of this review recommend that, in designing studies comparing aspects of questionnaire design, researchers should draw on theories of perception, pattern recognition and cognition[145] and seek to test the common recommendations of survey experts.[1,7,13,40] Comparative studies should use multiple outcome measures, including:

- the quantity of response (instrument response rates; item non-response rates)

Do you take a daily dose of aspirin?
(Please **circle the number** that describes you)

Yes, obtained on prescription 1

Yes, bought over the counter 2
(e.g. in a chemist or supermarket)

No, I don't take daily aspirin 3

FIGURE 1 *Example of vertical format for closed questions*

In the **past month**, on how many **days** have you been **short of breath during exercise** (for example going upstairs, walking up hill, gardening, taking part in sports)?

Never	On one or a few days	On several days	On most days	Every day
1	2	3	4	5

FIGURE 2 *Example of horizontal format for rating scale*

- the quality of response (non-response bias; validity, reliability and distribution of responses)
- resource implications (cost per completed questionnaire).

In addition to the quantifiable measures identified above, cognitive testing[134–137] of how respondents react to different design features should be employed.

Priorities for research

No evidence on the following aspects of questionnaire appearance was identified: double- versus single-sided printing; placement of questions within pages; print details; cover design; methods of identifying questions; vertical versus horizontal response formats; placement of headings and codes for response categories; identification of skip patterns; and nature, placing or format of instructions. However, the authors of this review regard some of these topics (e.g. double- versus single-sided printing) as of low priority for future research and present below their priorities for research, in order of importance.

- As noted in the recommendations for practice, computer-assisted surveys (OMR and OCR technology; web-based questionnaires) are likely to assume growing importance in the future. It seems likely that design principles for such questionnaires may differ from those espoused for paper-based questionnaires. Research into this topic would be very timely.
- Further testing of the impact of questionnaire length is desirable, particularly when the topic may be perceived as less salient.
- Formal testing of vertical versus horizontal response formats for multiple-choice questions is recommended.
- Studies on the relative placement of headings, response category descriptors and codes are advocated.
- Studies on verbal and graphical methods (including the use of colour contrast and different typefaces) to aid "navigation" through the questionnaire should be carried out.
- Studies on the placement and format of instructions for interviewers, respondents and data processors are required.

TABLE 9 *Length of questionnaire*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Criteria for comparison	Main findings
Cartwright, 1986[47]	RCT	3	Health: new mothers' experiences of pregnancy and labour	New mothers (UK)	Postal survey	Long: 24 pages, 110 questions (320) Medium: 16 pages, 65 questions (640) Short: 8 pages, 35 questions (640) (Also manipulated content of questionnaire (factual and attitudinal questions vs. factual only – short and medium versions)) Inclusion of blank sheet of paper for comments	Instrument response rates	Response rates: All content types: 78% long; 79% medium; 82% short Mixture of factual and attitudinal questions: 76% medium; 81% short Factual questions only: 82% medium; 84% short RR = 0.96 (95% CI, 0.91 to 1.02) medium vs. short (all content types) RR = 0.95 (95% CI, 0.89 to 1.02) long vs. short (all content types) RR = 0.99 (95% CI, 0.92 to 1.06) long vs. medium (all content types) RR = 0.93 (95% CI, 0.86 to 1.01) medium vs. short (facts + attitudes) RR = 0.97 (95% CI, 0.91 to 1.04) medium vs. short (facts only)
Jacoby, 1990[150]	RCT	4	Health: patients' experiences of and satisfaction with primary care services	Primary care patients (UK)	Postal survey	Long: 16 pages, 50 questions (1000) Short: 8 pages, 30 questions (1000) (Also manipulated sponsorship) See also *Table 23*	Instrument response rates	Response rates: Regardless of sender: 67.8% long; 68.4% short When sent by research organisation: 65% long; 65% short When sent by FPC: 79% long; 82% short RR = 0.97 (95% CI, 0.81 to 1.17) long vs. short (regardless of sender) RR = 1.00 (95% CI, 0.81 to 1.23) long vs. short (sent by research organisation) RR = 0.97 (95% CI, 0.81 to 1.17) long vs. short (sent by FPC)

continued

TABLE 9 contd *Length of questionnaire*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Criteria for comparison	Main findings
Hansen and Robinson, 1980[151]	RCT	4	Non-health: general public's attitudes towards most recent car purchase	Car purchasers (USA)	Postal survey	Long: 102 questions (300) Short: 32 questions (300) (Also manipulated nature of prenotification contact) See also *Table 17*	Instrument response rates Item non-response rates Speed of response	Response rates: Regardless of nature of prenotification: 32% long; 44% short No prenotification: 17% long; 30% short Yes/no "foot-in-door" prenotification: 33% long; 43% short Probe "foot-in-door" prenotification: 46% long; 58% short RR = 0.73 (95% CI, 0.59 to 0.90) long vs. short (regardless of prenotification) RR = 0.57 (95% CI, 0.33 to 0.96) long vs. short (no prenotification) RR = 0.77 (95% CI, 0.54 to 1.10) long vs. short (yes/no "foot-in-door") RR = 0.79 (95% CI, 0.61 to 1.04) long vs. short (probe "foot-in-door") Mean unanswered questions: No prenotification: 4.06% long; 4.02% short Yes/no "foot-in-door" prenotification: 4.01% long; 3.98% short Probe "foot-in-door" prenotification: 3.98% long; 4.01% short Response speed (mean days to respond): No prenotification: 14.2 long; 13.8 short Yes/no "foot-in-door" prenotification: 7.8 long; 7.5 short Probe "foot-in-door" prenotification: 7.7 long; 7.4 short (In ANOVA, main effect of length of questionnaire reported as "ns"; insufficient data to calculate CIs)
Layne and Thompson, 1981[152]	RCT	3	Non-health: scales measuring dogmatism, internality/ externality and assertiveness	University students (USA)	Postal survey	Long: 3 pages, 30 questions (200) Short: 1 page, 30 questions (200)	Instrument response rates	Overall response rate: 28% No significant association between response and questionnaire length $\chi^2 = 0.014$, df = 1, $p > 0.05$; insufficient data to calculate RR or CI)
Adams and Gale, 1982[153]	RCT	4	Non-health: participation in student body activities	University students (USA)	Postal survey	Long: 5 pages (550) Medium: 3 pages (550) Short: 1 page (550)	Instrument response rates	Response rates: 22% long; 47% medium; 41% short RR = 1.16 (95% CI, 1.02 to 1.33) medium vs. short RR = 0.55 (95% CI, 0.46 to 0.66) long vs. short RR = 0.47 (95% CI, 0.40 to 0.57) long vs. medium

continued

TABLE 9 contd *Length of questionnaire*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Criteria for comparison	Main findings
Powers and Alderman, 1982[154]	RCT	4	Non-health: opinions on scholastic aptitude test and related materials	School students (USA)	Postal survey	Long: 7 pages, 28 questions, 83 responses required (1004) Short: 6 pages, 20 questions, 69 responses required (1003) (Also manipulated offer of feedback) See also *Table 26*	Instrument response rates	Overall response rates: Regardless of feedback: 44% long; 52% short Feedback offered: 46% long; 54% short No feedback offered: 42% long; 49% short RR = 0.84 (95% CI, 0.77 to 0.92) long vs. short (regardless of feedback) RR = 0.84 (95% CI, 0.74 to 0.95) long vs. short (feedback offered) RR = 0.84 (95% CI, 0.73 to 0.96) long vs. short (no feedback offered)
Trice and Dolan, 1985[131]	RCT	3.5	Non-health: views of hotel services	Hotel guests (USA)	Self-completion survey	Study 1: Version 1: 10 items, no comments (200) Version 2: 5 items, open comments (200) Version 3: 5 items, structured comments (200) Study 2: Version 1: 15 items, open comments (200) Version 2: 10 items, open comments (200) Version 3: 5 items, open comments (200) (Also manipulated response categories) See also *Table 8*	Instrument response rates Rating of items Provision of additional comments	Response rates: Study 1: 12% version 1; 22% version 2; 24% version 3; 23% versions 2 and 3 combined Study 2: 26% version 1; 24% version 2; 20% version 3 Study 1: RR = 0.49 (95% CI, 0.31 to 0.77) version 1 vs. version 3 RR = 0.52 (95% CI, 0.33 to 0.83) version 1 vs. version 2 RR = 0.94 (95% CI, 0.65 to 1.34) version 2 vs. version 3 RR = 0.51 (95% CI, 0.33 to 0.77) version 1 vs. versions 2 and 3 combined Study 2: RR = 1.28 (95% CI, 0.89 to 1.84) version 1 vs. version 3 RR = 1.18 (95% CI, 0.81 to 1.71) version 2 vs. version 3 RR = 1.09 (95% CI, 0.77 to 1.53) version 1 vs. version 2 Mean ratings for common items (study 1): 20.9 version 1; 23.1 version 2; 23.4 version 3 (high score denotes a positive rating) (significance not reported and insufficient data to calculate CIs) Provision of additional comments: Study 1: 43% version 2; 24% version 3 Study 2: 25% version 1; 36% version 2; 48% version 3 Study 1: RR = 1.85 (95% CI, 0.99 to 3.42) version 2 vs. version 3 Study 2: RR = 0.54 (95% CI, 0.30 to 0.95) version 1 vs. version 3 RR = 0.76 (95% CI, 0.46 to 1.26) version 2 vs. version 3 RR = 0.92 (95% CI, 0.54 to 1.59 version 1 vs. version 2

continued

TABLE 9 contd *Length of questionnaire*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Criteria for comparison	Main findings
Trice, 1986[157]	RCT	4	Non-health: views of hotel services	Hotel guests (USA)	Self-completion survey	Version 1: 15 items, space for comments (150) Version 2: 15 items, no space for comments (150) Version 3: 5 items, space for comments (150) Version 4: 5 items, no space for comments (150) (Also manipulated: timing of completion (check-in vs. departure); incentives) See also *Table 25*	Instrument response rates	Response rates: 21% version 1; 20% version 2; 19% version 3; 14% version 4; 20% versions 1 and 2 (long) combined; 16% versions 3 and 4 (short) combined; 20% versions 1 and 3 (space for comments) combined; 17% versions 2 and 4 (no space for comments) combined RR = 1.23 (95% CI, 1.04 to 1.46) long vs. short (V1 + V2 vs. V3 + V4) RR = 1.08 (95% CI, 0.86 to 1.36) version 1 vs. version 3 RR = 1.44 (95% CI, 1.11 to 1.86) version 2 vs. version 4 RR = 1.19 (95% CI, 1.00 to 1.41) space for comments vs. no space for comments (V1 + V3 vs.V2 + V4)
Helgeson and Ursic, 1989[80]	RCT	3	Non-health: evaluation of product brands in fast food restaurants	University students (USA)	Self-completion survey	Long: 50 questions (60) Short: 10 questions (60) (Also manipulated mode of admin.) See also *Table 5*	Decision-making processes	Significant interaction ($p < 0.10$; insufficient data to calculate CIs) between mode of admin. and length of questionnaire for 3 types of decision-processing codes Proportion of decision-processing codes designated as "specific statements about past product usage experience" or as "neither cognitive nor affective" remained stable across questionnaire length for electronic data collection but dropped off for longer pencil-and-paper questionnaire Proportion of decision-processing codes designated as "statement of brand name" remained stable across questionnaire length for electronic data collection but increased (by 5%) for long vs. short pencil-and-paper questionnaire
Roszkowski and Bean, 1990[155]	RCT	2	Non-health: evaluation of distance learning courses	Students on distance learning course (USA)	Postal survey	Long version: 2 pages (5034) Short version: postcard, equivalent to ½ page (3500)	Instrument response rates Rating of items	Response rates (mean across 14 replications): 53% long; 81% short RR = 0.65 (95% CI, 0.63 to 0.67) long vs. short Respondents who were "very satisfied" (mean across 14 replications): 30% long; 28% short RR = 1.07 (95% CI, 0.98 to 1.17) long vs. short
Herzog and Bachman, 1981[156]	Non-random concurrent controlled study	3	Non-health: experiences and views on wide range of topics	School students (USA)	Self-completion (captive audience) survey	Long: 2 h + to complete (1050) Short: <45 min to complete (18924)	Pattern of responses	% using a single response category for all items in a series of related questions increased in later part of long version of questionnaire; this trend not apparent for short version Mean standardised differences between long and short versions larger in later part of questionnaire Correlations between adjacent items in series of questions more marked in later part of long questionnaire (Data not presented in sufficient detail to allow calculation of CIs)

FPC, Family Practitioner Committee

TABLE 10 *Pagination*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Criteria for comparison	Main findings
Suhre, 1989[163]	RCT	3	Non-health: views of innovation aspects of education and school effectiveness characteristics	School principals (The Netherlands)	Postal survey	A4 questionnaire (494 approx.) A5 questionnaire (494 approx.) Exact nos not stated in article; total sample size of 988 with "approximately equal numbers" receiving each treatment (Also manipulated incentives) See also *Table 25*	Instrument response rates	Overall response rate (mean): 52% No significant difference (p = 0.18) for format of questionnaire (insufficient data to calculate CIs)

TABLE 11 *Colour of paper*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Criteria for comparison	Main findings
Jobber and Sanderson, 1983[164]	RCT	3.5	Non-health: marketing strategies of textile companies	Directors of textile companies (UK)	Postal survey	White paper (400) Blue paper (400) (Also manipulated prenotification) See also *Table 17*	Instrument response rates	Response rates: Final: 56% white paper; 59% blue paper To 1st mailing: 39% white paper; 41% blue paper To follow-up: 28% white paper; 30% blue paper RR = 1.06 (95% CI, 0.94 to 1.19) blue vs. white (final response) RR = 1.07 (95% CI, 0.90 to 1.27) blue vs. white (1st mailing) RR = 1.08 (95% CI, 0.81 to 1.43) blue vs. white (follow-up) Non-significant interaction (p > 0.05) between colour of paper and whether a prior letter sent (insufficient data to calculate CI) No significant difference (p > 0.05) in response rates with and without follow-up between white and blue paper (insufficient data to calculate CI)

TABLE 12 *Response format*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Criteria for comparison	Main findings
Cartwright, 1986[47]	RCT	4	Health: new mothers' experiences of pregnancy and labour	New mothers (UK)	Postal survey	"Circle the number" (300) "Tick the box" (300)	Instrument response rate Item non-response rate	Response rates: 78% "circle the number"; 80% "tick the box" RR = 1.03 (95% CI, 0.94 to 1.11) "tick the box" vs. "circle the number" Average no. inadequate answers (to 94 precoded questions) per questionnaire (overall): 3.17 "circle the number"; 2.82 "tick the box" (p = ns; insufficient data to calculate CI) Average no. inadequate answers per completed questionnaire (adjusted for difference in completion rates): 1.27 "circle the number"; 1.10 "tick the box" (p = ns; insufficient data to calculate CI)

Chapter 6

Enhancing response rates

Introduction

The importance of high response rates

The confidence with which findings from a survey can be accepted and generalised is a function, in part at least, of achieved sample size, which is directly related to response rate. Surveys in which the achieved sample is small have low precision; in other words, the CIs around any estimates of population parameters (e.g. mean age, percentage holding a particular opinion) are wide. The generalisability of findings also depends on how representative the sample is of the underlying population. Poor response rates are a potential source of bias (systematic error) because non-respondents are likely to differ from respondents with respect to important characteristics. Sackett,[4] citing Cochran and Chambers[166] and Murphy,[23] suggested that, in both case-control studies and cohort studies, the effect of non-respondent bias on estimates of relative odds can be either to increase or decrease such odds. Moser and Kalton[5] demonstrated mathematically how, in estimating the mean for an entire population, the magnitude of non-response bias is a function of both the percentage of the sample who do not respond and the extent to which the population mean for the non-respondents differs from that for the respondents.

To reduce the threat of non-response bias and to increase the precision of estimates, survey researchers should devote effort to enhancing response rates. However, they should also take care that their efforts do not lead to response bias[149] or to sample composition bias. Response bias occurs when the answers provided by respondents are invalid; in other words, they do not reflect their true experience or attitudes. Sample composition bias occurs when an inducement technique acts selectively, leading to systematic differences between the achieved sample and the underlying population. In practice, without a standard against which to validate data from the sample, it may be difficult to separate the two effects.

Experts differ in their views on what constitutes an adequate response rate. Fowler[14] recommended that the standard for a minimum acceptable response rate should be set at 75%. Mangione[16] indicated that, for postal surveys, response rates in excess of 85% are "excellent"; those in the range 70–84% are "very good"; 60–69% are "acceptable"; 50–59% are "barely acceptable"; and response rates below 50% are "unacceptable". Borg and Gall[167] suggested that, if the rate of non-response is in excess of 20%, the findings of the study are likely to be altered (in other words, bias is likely to occur).

By these standards, response rates reported in the medical literature are low. Asch and colleagues[168] found that the mean response rate among surveys conducted in the USA and reported in American medical journals was only 60%; for published surveys of physicians the rate was even lower at 54%.

Sources of non-response

The most crude measure of response rate[169] (or, as it is sometimes termed, "survey completion rate") is "the total sample minus all uncompleted interviews (regardless of cause) divided by the total sample". Clearly, "questionnaires" can be substituted for "interviews" in the case of self-completion surveys. However, this definition assumes a perfect sampling frame. In reality, sampling units outside the study population (e.g. people who have died or do not meet study eligibility criteria) should be "considered as blanks on the sampling frame" rather than as non-respondents.[5] These blanks or ineligibles should be removed from the denominator and the response rate calculated as completed interviews or questionnaires divided by this revised denominator.[170] If feasible, it is reasonable to replace sampling units recognised to be ineligible by substitutes, selected appropriately from the same sampling frame.

The five main sources of non-response that tend to occur in a survey are:[5]

- "Unsuitable for inclusion"; for example, those who are too infirm, deaf, blind, illiterate or non-competent in the language of the survey: Obviously, deafness is more of a hindrance in interview surveys, whereas blindness and illiteracy have a greater impact on self-completion questionnaires. Moreover, with care, the researcher can minimise the threat of non-response of this type, for example, by providing translated versions of the questionnaire or

using interviewers who are competent in the appropriate language, or interpreters.
- Movers: In any preselected sample, it is likely that a number of those sampled will no longer live at the listed address and will be untraceable to their new address. Some movers may in fact become ineligible by moving outside the study area. The threat of non-response of this type is related to how up to date the sampling frame is; the techniques described in the remainder of this chapter are not aimed at reducing this form of non-response.
- Refusals: Inevitably, some of those sampled will simply decline the invitation to participate in the survey. In interview surveys, refusals are generally explicit; in postal surveys, a significant proportion of refusals may be implicit (i.e. the individual may simply never return the questionnaire). In interview surveys, good contacting and "door-stepping" techniques[51,171] can reduce this source of non-response.
- Away from home: In a time-limited survey, some sampled individuals may simply be absent from home for the entire duration of the survey.
- Not-at-homes: This source of non-response is a much greater threat in interview surveys (both face-to-face and telephone interviews). With well-trained interviewers and clearly defined contact protocols, it may be possible to convert some of these non-respondents to respondents by repeated contacts on different days or at different times. In postal surveys, not-at-homes are unlikely to be a source of non-response unless intensive mailing techniques, requiring the addressee to sign for the delivery, are used.

In postal surveys it may be difficult to estimate accurately the relative contributions of each type of non-response; for example, non-contact is not always confirmed by the postal service[172] and may be misclassified as refusal. In the UK, the Royal Mail will return undeliverable mail to the sender only if the sender's address is displayed on the outgoing envelope (preferably on the back).

Characteristics of non-respondents

It is generally held that non-respondents, in both interview surveys and postal surveys, tend to differ in important ways from the underlying population. This view is supported by empirical evidence[174,175] showing that respondents to postal surveys, particularly those returning their questionnaires early, are likely to be more interested in the survey topic, to make more favourable reports and to be more successful in their current status. However, evidence of other forms of non-response bias is less clear-cut. Kanuk and Berenson,[159] in an extensive literature review of factors influencing response rates to postal surveys, examined the differences between respondents and non-respondents with respect to a wide range of demographic, socio-economic and personality characteristics. They concluded that the only consistent and widespread finding was that respondents tend to be better educated and therefore to have greater facility in writing. In respect of interview surveys, Goyder[175] demonstrated that survey response rates (given that initial contact has been made) tend to be positively correlated with socio-economic status and negatively correlated with age.

In health surveys of the general population and of specific patient groups, non-respondents to postal surveys are more likely to be in semiskilled or unskilled manual occupations and to be from ethnic minorities, while respondents have been shown to be more likely to be younger, have higher levels of educational attainment, and have better health status;[176] this latter finding is of particular relevance when the aim is to measure health status in the underlying population because estimates derived from survey respondents are likely to be upwardly biased.

Non-response bias also occurs in postal surveys of professional groups. Cartwright,[177] in an overview of 19 surveys of health professionals (including both interview and postal surveys), reported that relatively higher response rates were generally obtained in surveys of nurses. Younger doctors and those with better qualifications, and consultants with university rather than NHS appointments, were more likely to respond. Single-handed GPs were less likely to respond. Non-responding doctors were somewhat less likely to be regarded as sympathetic and helpful by their patients. However, response rates for GPs and consultants were broadly similar, and the gender of doctors did not generally have a significant impact on response rates. Cartwright[177] also concluded that the number and direction of biases did not appear to be strongly related to overall response rate, but of concern was her observation that response rates amongst doctors appeared to be dropping in recent years, a finding echoed by McAvoy and Kaner.[49]

It has also been suggested[5] that late responders to postal surveys may resemble non-respondents more closely than they resemble early respondents. If this is indeed the case, it may be possible to infer something about the characteristics and attitudes of non-respondents from data gathered from late respondents. However, findings of differences between early and late respondents are by no means

universal. For example, Hovland and colleagues,[178] in a survey of dentists, found no significant differences in attitudes or knowledge scores between those who responded to the first mailing and those who replied only after reminders, although Fiset and colleagues,[179] in a separate survey of dentists, found differences between early and late respondents with respect to demographic characteristics and dissatisfaction with dental practice.

Theoretical perspectives on enhancing response rates

Appropriate methods for enhancing response rates depend on the likely sources of non-response. For example, if a significant number of sampled individuals are anticipated to be non-fluent in the original language of the survey, it may be appropriate to translate the questionnaire. Techniques for dealing with movers, especially in interview surveys, are described in some detail by Moser and Kalton.[5] Theories from the fields of sociology and psychology provide a useful insight into respondent behaviour and the likely impact of different inducements, particularly in reducing non-response due to refusals.

The theory of social exchange[142,143] suggests that the actions of individuals are influenced by the rewards they expect to obtain from completing these actions and by the costs of doing so. If individuals are behaving rationally, they will seek to keep the costs of an activity below the rewards they expect to derive from its performance. Thus respondents will respond to a survey only if the anticipated rewards of participation are at least equal to or exceed the costs of responding. This implies that there are three things that the surveyor must do to maximise response rates:[1]

- minimise the costs (physical, mental, emotional and economic) of responding
- maximise the rewards (tangible and intangible) for responding
- establish trust that those rewards will be delivered.

Table 13 presents examples of ways in which each of these objectives can be achieved.

TABLE 13 *Experts' recommended means of achieving objectives in maximising response rates (after Dillman[1])*

Minimising cost of responding	Maximising rewards of responding	Establishing trust
Making questionnaire clear and concise Attention to issues of question wording and sequencing	***Making questionnaire interesting to respondent*** Choice of topic Addition of "interesting" questions	***Establishment of benefit of participation*** Statement of how results will be used to benefit respondents/others Promise to send results of research
Making questionnaire (appear) to be simple to complete Attention to issues of questionnaire appearance	***Expression of positive regard for respondent as an individual*** Stating importance of individual's contribution Personalised salutation Individually typed letter Handwritten signature Stamped (not franked) mail	***Establishment of credentials of researchers*** Use of headed notepaper Naming of researchers
Reduction of mental/physical effort required for completion and of feelings of anxiety/inadequacy Simple questions Clear instructions Sensitive handling of potentially embarrassing questions	***Expression of verbal appreciation*** Statement of thanks in all communications Statement of thanks on questionnaire Follow-up "thank you" letter or card	***Building on other exchange relationships*** Endorsement by well-regarded organisation/individual
Avoidance of subordination of respondent to researcher	***Support of respondent's values*** Appeal to personal utility Appeal to altruism/social utility	
Reduction of direct monetary costs of responding Provision of prepaid return envelopes	***Incentives*** Monetary or material incentive at time of response Provision of results of research	

Dillman[1] noted that there are important differences with respect to the application of social exchange theory between postal and telephone surveys. Although he did not consider face-to-face interviews, the relative importance of the various exchange considerations for that mode of survey administration is also likely to be different. The remainder of this chapter will concentrate on methods for enhancing response rates in postal and other self-completion surveys. In interview surveys, the social interaction between interviewer and interviewee means that non-response is most likely to occur before the interview proper has been initiated. Therefore, contact approaches and "door-stepping" techniques[5,51,171] are likely to be the most potent strategies, although some of the factors influencing response rates to self-completion questionnaires – prenotification, saliency, sponsorship, incentives and the offer of feedback – are also likely to come into play.

Brown and colleagues[140] developed the framework presented in *Table 13* into a four-stage task-analysis model of decision-making. The decision to proceed to each successive stage is influenced by many factors, including aspects of questionnaire design and survey administration, as summarised in *Table 14*.

- Stage 1: Evocation of interest in the survey. Unless recipients' interest is aroused, they will not give any further consideration to the task.
- Stage 2: Evaluation of the task of participation in the survey. If the costs of participation are perceived to outweigh the benefits, the recipient will proceed no further.
- Stage 3: Initiation and monitoring of the task of completion, which may be abandoned at any time if it is perceived to be too burdensome.
- Stage 4: Decision to return the completed questionnaire.

TABLE 14 *A task-analysis model of respondent decision-making (after Brown* et al.*[140])*

Stage 1 **Interest in task**	**Stage 2** **Evaluation of task**	**Stage 3** **Initiation and monitoring of task**	**Stage 4** **Completion of task**
Personal contact Personalisation of letter Personalisation of envelope Class of mail	***Time and effort required*** Length of questionnaire Size of pages Supply of addressed return envelope Supply of stamped return envelope	***Actual difficulty encountered*** Clarity of question wording Clarity of instructions Complexity of questions	***Provision of stamped addressed envelope***
Questionnaire appearance Cover illustration Colour of cover Layout and format Quality/clarity of type	***Cursory evaluation of difficulty*** Number of questions Complexity of questions	***Sensitivity of requests*** Number and nature of sensitive questions	***Reminders to return***
Topic Questionnaire title Cover illustration Content of covering letter Timeliness Relevance/salience		***Actual time required***	
Source credibility/trust Image of sponsor Credentials of individual investigator Message in covering letter			
Reward for participation Tangible rewards: monetary and other incentives Intangible rewards: appeals to altruism, self-interest etc.			
Persistence of source Follow-up procedures			

Sudman[180] suggested four possible reasons for non-response from professional groups: potential participants are too busy; the value of the survey is either not apparent or is perceived to be low; potential participants have concerns about confidentiality; and they have concerns about the validity of the questions. He argued that professionals "cannot simply be treated as members of the general population but must receive incentives and information not usually necessary for the general population", a point echoed by Ward[181] in relation to encouraging GPs to participate in research. Sudman[180] also recognised the existence of "hard-core" refusers (those who perceive that postal surveys are intrinsically invalid) and suggested that this group should be offered the opportunity to identify themselves and opt out after the initial mailing.

Identification of primary studies

Methods of enhancing response rates are perhaps the most researched aspects of survey design and administration. In 1990, Gajraj and colleagues[182] stated that "a recently published bibliography on marketing research methods cites 454 studies of mail survey responses" (p. 42). The cited reference had been published in 1986.[27] An ongoing systematic review of methods to influence response rates to postal questionnaires has identified 282 published randomised controlled trials (Edwards P, Institute of Child Health, University College, London: personal communication, 2001).

In this review, emphasis is on techniques for enhancing response rates in postal surveys. A total of 68 randomised controlled trials and three quasi-experimental studies were identified (mainly involving systematic allocation to intervention and control groups) in which factors hypothesised to influence response rates were manipulated. None reported an explicit power calculation to determine sample size. In assessing quality in respect of the ability to calculate RRs and CIs, the focus was solely on the primary outcomes of response rates; almost all studies reported findings in sufficient detail to permit these calculations. In many of the identified studies, multiple factors were manipulated simultaneously, generally using a factorial design. In such studies, the original researchers usually looked for both the main effects of each factor and for any interactions between factors; even when they did not formally analyse all main effects, results were usually presented in sufficient detail to allow complete analyses. Multifactorial studies are reported below and in the accompanying tables under all the appropriate headings. However, in the interest of parsimony, the multiple factors are not repetitively described in the text; in the tables, only the main effect under scrutiny is reported.

The specific issues examined were:

- mechanical and perceptual factors, further subdivided into:
  - timing of survey
  - number, timing and method of contacts
  - postage rates and types
- general motivational factors, further subdivided into:
  - anonymity/confidentiality
  - personalisation
  - nature of appeal; other aspects of covering letter
  - sponsorship
  - saliency
- financial and other incentives
- miscellaneous factors.

Other factors influencing response rates are discussed elsewhere in this report. The effects of question wording and ordering are discussed in chapter 4, while the impact of questionnaire length and format is considered in chapter 5.

Tables *15–27* (see pp. 135–174) identify the primary and secondary outcomes. Primary outcomes are in all cases some type of instrument (questionnaire) response rate; secondary outcomes include speed and cost of response, response quality, response bias and sample composition bias. RRs and corresponding 95% CIs for response rates are presented in the tables. Significant findings with respect to secondary outcomes are highlighted in the text.

Some caveats are needed in examining these findings from primary research. Many of the identified studies involved surveys of the general public, among whom response rates are typically significantly lower than for surveys of special populations such as patients or professional groups.[160] Furthermore, the low response rates achieved in many of the identified studies may be attributable in part to the apparently low saliency of the survey topics. Heberlein and Baumgartner's review[160] showed that saliency (a salient topic being defined as "one which dealt with important behaviour or interests that were also current") is one of the strongest predictors of response rate. Health-related topics are generally considered to have greater saliency.

In interpreting and comparing the evidence from the identified primary studies, it is important to be aware that factors other than those that are experimentally manipulated (e.g. intensity of follow-up) varied from study to study. This heterogeneity makes it difficult to compare response rates between studies.

Timing of survey

Dillman[1] recommended a mailing date early in the week; he suggested that mailing on a Monday or a Tuesday allows questionnaires to be forwarded to a new address for receipt in the same week. He also recommended avoiding the month of December, on the basis of competing pressures on people's time. The review by Brown and colleagues[140] drew on findings from surveys of leisure activities carried out at different times of the year and found that January, February and March were the optimal months for survey research, at least on leisure topics.

Identified studies

Only one study (*Table 15*; see p. 135) examining the effects of timing of delivery that met the criteria was identified. This was a randomised trial on a health-related topic. In a survey of doctors, Olivarius and Andreasen[183] investigated the impact on response rates of posting a questionnaire on Thursday (to arrive on Friday) or on Saturday (to arrive on Monday); they conjectured that receipt just before a weekend, when doctors theoretically have more free time, would lead to an improved response rate. However, no effect of day of posting was found either for GPs or for specialists.

Number and relative timing of contacts

In their review, Heberlein and Baumgartner[160] showed that the total number of contacts (including both prenotification and follow-up) had the strongest zero-order correlation with response rates ($r = 0.634$; $p < 0.001$), accounting for 42% of observed variance in response rates. In their simple regression model, each additional contact increased the predicted response rate by 12%. Dillman's Total Design Method[1] involves four contacts in total: the initial mailing and up to three follow-ups.

Identified studies

Two studies were identified[110,184] (*Table 16*; see p. 135), both of which focused on:

- the number of contacts
- the relative timing (i.e. prenotification versus follow-up) of contacts.

Both were randomised trials but neither was on a health-related topic. Both showed a significant impact of multiple contacts on response rates, and showed that postnotification (i.e. follow-up/reminders) was more powerful than prenotification in stimulating response.

Jones and Lang,[110] in a survey of house purchasers, compared prenotification, postnotification (essentially a reminder), and both pre- and postnotification. For those receiving a single contact only, postnotification was more effective than prenotification in stimulating response. Furthermore, as hypothesised, response rates were significantly higher for multiple contacts. However, the hypothesis that multiple contacts would be better than a single contact in reducing sample composition bias was not supported. For house purchase price, the distributions under both prenotification alone and under pre- plus postnotification were both statistically identical to the underlying population distribution, while number and timing of contacts had no observable effect on the distribution of house purchase dates within the achieved sample. Under the postnotification condition, bias in favour of purchasers of more expensive houses was observed, although no bias with respect to date of purchase was found. Contrary to expectations, the number and timing of contacts alone did not lead to significant response bias (i.e. mismatch between reported and actual house price or purchase date). However, there was a significant (and difficult to interpret) interaction between timing of contact and questionnaire format. A combination of prenotification and "attributes questions first" format led to negative bias with respect to reporting date of purchase (i.e. reporting the purchase date as less recent than it really was).

In a very large experiment (10,800 respondents), Peterson and colleagues[184] manipulated the number and relative timing of contacts (between one and four contacts from: advance notification; initial mailing of questionnaire; first reminder; second reminder), mode of contact (reminders in the form of a letter or postcard), as well as personalisation, spon-sorship and saliency. There was a significant linear trend in response rates with respect to number of contacts; each additional contact resulted in an increase of approximately 4% in response rates. Postnotification (one or two reminders) was more effective than prenotification in stimulating response. The strategy producing the highest response rate involved four contacts in total:

prenotification by letter, mailing of initial questionnaire, first reminder by postcard, and second reminder by letter with a duplicate questionnaire. The response rate from this strategy was 28% compared with 10% for a single contact (initial mailing of questionnaire only). The net costs per response for these two extremes were $7 and $6 respectively. Average costs per response were: $6.27 for a single contact; $8.03 for two contacts; $8.56 for three contacts; and $8.82 for four contacts. When all 27 possible strategies were compared, there was no systematic association between response rate and response cost ($r = 0.02$), suggesting that the selection of an optimal strategy depends on whether the goal is maximisation of response rate, minimisation of cost, or some balance between the two.

Prenotification contacts

Childers and Skinner,[185] drawing on the exchange paradigm used by Bagozzi[186] to conceptualise marketing behaviour, recognised that stimulating response to postal surveys represented a special "marketing" problem, attributable in part to the "limited nature of prior and subsequent contact" (p. 40). Making an approach to target respondents prior to sending the questionnaire may be one way of addressing this problem.

A particular form of prenotification is the active "foot-in-door" approach.[187] This technique involves gaining the respondent's compliance with a small request (the "foot", e.g. answering a few initial questions) with the ultimate goal of gaining compliance with a larger request. The theory behind this approach is that individuals develop perceptions or attributions of themselves based upon observations of their own behaviour and the situational context in which this behaviour occurs.[188] Thus, if someone complies with a small request, he or she is likely to develop a self-perception of being a cooperative person, enhancing the probability of acceding to future requests for compliance with larger tasks.

Linsky's review[158] concluded that prenotification appears to increase response rates and is particularly effective if it is in the form of a telephone call. However, in their later review, Heberlein and Baumgartner[160] believed that this zero-order effect disappeared on controlling for total number of contacts. They found prenotification to be no more or no less effective than follow-up contacts.

Identified studies

Eight articles[151,164,184,187,189–193] investigated aspects of prenotification (*Table 17*; see p. 137). All were randomised trials but only one was on a health-related topic.[187]

Comparisons involved:

- any prenotification versus no prenotification: four studies[164,184,191–193] (Martin and colleagues reported the findings from the same study in two separate articles)
- different modes of prenotification: three studies[184,190,193]
- different "foot-in-door" techniques with each other and with no prenotification: three studies.[151,187,189]

Prenotification versus no prenotification

Of the four studies that examined the impact of a prenotification contact, two[184,191,192] reported a wholly positive effect on response rates.

Martin and colleagues[191,192] looked at the effects of prenotification, follow-up, personalisation and type of postage, using a factorial design. Prenotification was found to lead to significantly higher response rates. There was a significant interaction between prenotification and personalisation. The additional cost for sending an advance letter was $0.53 per questionnaire sent, yielding a cost for each additional questionnaire returned as a result of prenotification of $3.33.

Peterson and co-workers[184] conducted a complex factorial experiment in which the number and nature of contacts, personalisation of address label, survey sponsorship and saliency were also manipulated. Prenotification was either by letter or by postcard. Regardless of the mode of contact, prenotification led to a significantly higher response rate (although overall response rates in this survey were low, averaging only 18%).

Findings from the study by Faria and colleagues[193] were mixed. Comparing any mode of precontact with no prenotification indicated that prenotification significantly enhanced response rates. However, decomposition by mode of precontact showed a mode effect: prenotification by letter led to a significantly higher response rate than no precontact, but the difference between telephone prenotification and no precontact did not reach statistical significance (in contrast to the findings from Linsky's review[158]).

Finally, in the study by Jobber and Sanderson,[164] and contrary to the expectations of the researchers themselves, a higher (although not statistically significant) response rate was obtained from the

group of respondents who were not originally prenotified; this difference was accounted for mainly by responses received after a reminder.

Mode of prenotification

None of the three identified studies showed a significant effect of mode of contact on response rates.

Nederhof[190] compared prenotification by mail and by telephone. Although the mode of prenotification made no significant difference to response rates, a telephone contact led to sample composition bias; compared with mail prenotification, it increased response rates for men and decreased them for women.

Although, as already noted, Peterson and colleagues[184] found that prenotification *per se* led to a significant increase in response rates, there was no significant effect of mode of prenotification (letter versus postcard).

Faria and colleagues[193] compared two methods of precontact (telephone and letter) with no prenotification. Prenotification by letter led to a significantly higher response rate than no precontact, but the differences between telephone prenotification and no precontact, and between telephone and mail precontact, did not reach statistical significance. Although the trend in response speeds was in favour of prenotification, differences in days to respond (mean 8.67 days for no precontact; 8.36 for precontact by mail; 8.28 for precontact by telephone), and in the percentage responding within 1 week of initial mailing, did not reach statistical significance. Quality of response, measured in terms of item omission, was uniformly high (mean number of items omitted: 0.19 for no precontact; 0.14 for mail precontact; 0.28 for telephone precontact). Cost per response was lowest when no prenotification took place ($2.09 for no precontact; $2.52 for mail precontact; $3.06 for phone precontact).

Content of prenotification message ("foot" techniques)

All three studies of "foot" techniques involved control groups with no prenotification and they all showed that prenotification increased response rates relative to the control group, but findings regarding the effectiveness of different introductory messages were mixed.

In the only health-related study, Kamins[187] examined a range of "foot-in-door" approaches. The sample was divided into five groups. The first received no prenotification of the survey, while members of the second were contacted by telephone and asked if they would be willing to participate in a subsequent mail survey (solicitation control approach). The third group received a simple "foot-in-door" approach; in a telephone call, they were asked to indicate agreement/disagreement with four simple questions on healthcare, before being invited to participate in the mail survey. The fourth group received a "probe" foot-in-door approach; after each of the four agree/disagree questions, the interviewer probed for elaboration of the reasons for the response given. It was hypothesised that probing would increase involvement in the topic and therefore the likelihood of responding to the mail survey. Those in the fifth group were subjected to a "labelled probe" foot-in-door approach; in addition to the probing described above, they were told that they were "co-operative and helpful", both after answering the four questions and after agreeing to participate in the mail survey. It was hypothesised that this would enhance these individuals' self-perceptions of helpfulness, thereby increasing their propensity to respond. Analysis showed that there were significant differences in response rates to the initial mailing, to the follow-up mailing and overall between the different prenotification approaches. All forms of prenotification combined significantly increased response rates; however, the differences between the solicitation call and no prenotification, and the simple foot and no prenotification did not reach statistical significance. The labelled probe outperformed all other approaches with respect to initial response rates and was significantly better than the simple foot, solicitation and control groups with respect to final response rates; the probe foot was superior to the solicitation and no prenotification groups with respect to final response rates (although the 95% CI for the former comparison includes unity). No evidence of differential patterns of response was found when responses to attitudinal and importance questions were compared across the five approaches, suggesting that no response bias had occurred.

Allen and colleagues[189] compared a simple solicitation prior telephone call (in which potential respondents were told about the forthcoming mail survey and asked if they would participate), and a foot-in-door prior call (in which three introductory questions on issues related to the survey were posed before information on the mail survey was given) with the traditional mail survey approach of no prenotification. Of the 239 persons contacted by telephone, 196 (82%) agreed to take part in the mail survey; 56% of these (46% of all those contacted) subsequently returned a completed

questionnaire. The response rate among those who were prenotified was significantly higher (69% for simple solicitation; 67% for the foot-in-door approach) than in the group receiving no prior notification (22%). However, contrary to expectations, the questioning foot-in-door approach did not yield a significantly higher response than the simple solicitation.

In a similar experiment, Hansen and Robinson[151] compared low and high involvement foot-in-door approaches with no prenotification. In both the low and high involvement approaches, target respondents were telephoned and asked to answer some basic questions on the topic of the mail survey; in the "low involvement" group (yes/no prior), the interviewer simply asked whether the respondent agreed with a given statement, while in the high involvement approach (probe prior), the interviewer probed the reasons for their views. All those contacted were sent a questionnaire in the post within 3 days of the initial contact. It was hypothesised that any form of prenotification would lead to enhanced response rates, but that the probing format would be significantly better than the yes/no format. The findings supported these hypotheses. However, although 94.5% of those receiving a probe format call had agreed to complete a questionnaire, only 52% in fact did so; the corresponding figures for the yes/no format call were 86% and 38%. As hypothesised, the speed of response was greater for respondents who had received some form of prior notification (mean days to respond: 14.0 for no prior call; 7.65 for yes/no prior call; 7.55 for probe prior call). However, the hypothesis of greater response completeness after prenotification was not supported (item non-response rates averaged 4% regardless of presence or form of prenotification). In this study, the length of the questionnaire was also varied; the findings persisted on controlling for length.

Follow-up contacts (reminders)

Linsky's[158] review showed that follow-up contacts are generally effective in stimulating additional response, with intensive follow-ups (e.g. by telephone or special delivery mail) being of particular value; he suggested that, in the interest of keeping costs down, these intensive techniques should be reserved for persistent non-respondents.

A number of authors have demonstrated differences between early and later responders, suggesting that including only those who respond to an initial approach may introduce response bias. Moser and Kalton[5] suggested that the quality of response may decline in successive "waves" because those who are persuaded to reply by reminders may be less interested and may therefore take less care in answering. Sample composition bias may also occur. Fiset and colleagues,[179] in a survey of dental malpractice, showed that early respondents (i.e. those returning their questionnaire before any reminder was sent) tended to be older, male and Caucasian.

The use of reminders is generally endorsed in texts on survey methods.[5,13,40,194] Dillman's[1] Total Design Method involves three follow-up contacts. The first is a postcard, sent 1 week after the initial mailing of the questionnaire to all sample members, thanking those who have returned their questionnaire and reminding those who have not to do so. The second is a letter, enclosing a replacement questionnaire, sent to non-respondents 3 weeks after the initial questionnaire. The third is again a letter and replacement questionnaire, sent to persistent non-respondents, by certified mail 7 weeks after the initial questionnaire.

Identified studies

Perhaps surprisingly, only nine studies[184,191,192,195–201] were identified in which aspects of follow-up in mail surveys were manipulated in an experiment (*Table 18*; see p. 140). The conventional wisdom regarding the effectiveness of reminders in stimulating response appears to have been derived largely from comparisons of initial and final response rates within studies (i.e. before and after reminders), and of differential response rates between surveys in which follow-up contacts were and were not made. All of the studies identified for this review were randomised controlled trials; two were on health-related topics.[198,201]

Comparisons involved:

- reminders versus no reminders (two studies)[184,191,192]
- number of reminders (one study)[184]
- content of the reminder message (three studies)[195,198,199]
- mode of contact (four studies)[184,196,200,201]
- inclusion of a duplicate questionnaire (three studies).[184,197,201]

Reminders versus no reminders

In both identified studies, a small but significant increase in response rates was achieved through the use of reminders.

In the first, which involved a complex factorial design, the researchers[184] compared zero, one and two reminders. In comparison with no reminders, sending at least one reminder led to a significant increase in response rates.

Martin and colleagues[191,192] used a factorial design study to assess the effects of follow-up, prenotification, personalisation and type of return envelope. On controlling for these other factors, a higher response rate was found in the group to whom a reminder was sent. The additional cost for sending a reminder was $0.55 per questionnaire sent, yielding a cost for each additional questionnaire returned as a result of follow-up of $13.41.

Number of reminders

Peterson and colleagues,[184] in their complex factorial experiment, showed that response rates after two reminders were significantly higher than those for a single reminder.

Content of reminder message

All three of the identified studies examined the effectiveness of indicating that further follow-up contacts would be made unless a completed questionnaire was returned. The findings were equivocal.

In a health-related study, Blass and colleagues,[198] building on a theory of adherence to group norms, investigated the effect of informing non-respondents that a large number of their peers had already returned a completed questionnaire. Like Nevin and Ford,[195] they also investigated the impact of an implied threat of a further follow-up approach if the questionnaire was not returned. Initial non-respondents were randomly assigned to receiving follow-up letters reflecting one of four conditions: consensus–threat; consensus–no threat; no consensus–threat; and no consensus–no threat. No significant differences in response rates between the conditions were observed. However, there was a significant consensus by threat interaction in respect of speed of response; participants responded faster when at least one factor was manipulated but a combination of the two factors had a weaker effect than either individual manipulation (mean days to respond: 5.94 for consensus–no threat; 6.35 for no consensus–threat; 7.49 for consensus–threat; 10.21 for no consensus–no threat).

Nevin and Ford[195] compared a follow-up letter with a "veiled threat" (i.e. a comment referring to the fact that the respondent had not, according to the researchers' records, replied to the initial questionnaire, which was intended to imply a threat of continued follow-up until the respondent returned a completed questionnaire) and a simple follow-up letter. The response rate to the veiled threat approach was significantly higher and a significantly higher percentage of responses were obtained within 5 days of the follow-up mailing. However, the type of follow-up letter did not have any significant impact on item non-response rates or on response bias, measured in terms of how the questions were answered.

Dommeyer[199] compared six different follow-up letters. Four had a "negative" appeal in the sense of threatening an interview follow-up if the questionnaire was not returned (threat 1: telephone interview on specified day; threat 2: telephone interview during specified fortnight; threat 3: face-to-face interview on specified day; threat 4: face-to-face interview during specified fortnight); the fifth was a casual appeal asking for return of the questionnaire by a specified date; the sixth enclosed an incentive (25 cents) for return of the questionnaire. Analysis showed that at least two groups had significantly different response rates and highlighted the group receiving the threat of a face-to-face interview on a specific day and the group receiving the incentive as the source of these differences (12% versus 24%). A comparison of the four negative versus the two positive appeals showed a difference in response rates (17% versus 21%) that approached statistical significance (RR = 1.28; 95% CI, 0.99 to 1.66). No significant differences with respect to response speed, response quality, item non-response rates or response bias were observed. Costs per usable questionnaire were: $1.54 for a casual appeal; $1.83, $1.43, $2.43 and $1.56 for threats 1–4 respectively; and $2.33 for 25 cents incentive.

Mode of contact

Heberlein and Baumgartner's[160] review indicated that the use of a special mailing technique (e.g. certified mail or special delivery), or a personal or telephone contact, increased response rates over a standard postal reminder. However, findings from the two studies that looked at special mailing techniques were contradictory.

Gitelson and Drogin[200] tested the effectiveness of certified mailing and personalisation in a third (final) follow-up; response rates prior to this follow-up were 67%. Certified mailing (which was always combined with personalisation) led to a significantly higher response rate (43% versus 17% for personalised standard mail and 13% for non-personalised standard mail), bringing overall response rates in the certified mailing group to

over 80%. They also examined whether sample composition bias was reduced by this final follow-up (regardless of mode); they found that the final mailing increased responses from individuals with less involvement in the survey topic, but differences with respect to behaviour and expenditure were not of practical significance. Finally, they noted that using certified mailing to contact all 450 individuals who had not responded to the first three mailings would have cost an additional $405.

Kahle and Sales[196] compared sending a final (second reminder) replacement questionnaire by certified mail or by first-class mail, with airmail stickers affixed. No significant effect of postage rate on response rates was noted.

The remaining two studies compared postcard and letter reminders, with mixed results. Peterson and co-workers[184] used a complex design that enabled the comparison of postcards and letters (including a duplicate questionnaire) for both first and second reminders. For first reminders, the difference in response rates between the two modes of contact was not statistically significant. However, for the second reminder, a letter and duplicate questionnaire significantly enhanced response rates over a postcard reminder.

Likewise, and in a health-related survey, Roberts and colleagues[201] found no significant differences in response rates to a postcard or letter reminder (this first reminder was sent 3 weeks after the initial questionnaire; the letter included a duplicate questionnaire and a freepost return envelope). In this study, after a further 3 weeks, all outstanding non-respondents were sent a second reminder in the form of a letter with a duplicate questionnaire and a freepost return envelope. Although response rates to this second reminder were significantly higher for those who had initially had a postcard reminder, the overall difference in response rates to the two reminders did not reach statistical significance. However, the use of a letter as a first reminder led to an overall cost per response that was 1.3 times that of the postcard (£2.77 versus £2.13). The authors concluded that a postcard as a first reminder is a "practical and economic strategy by which to increase response". Note, however, that these findings and conclusions may have been confounded by the enclosure of the duplicate questionnaire with the first letter.

Inclusion of duplicate questionnaire

Dillman[1] strongly advocated the inclusion of a duplicate questionnaire in follow-up mailings, on the grounds that non-respondents were likely to have mislaid the original. Heberlein and Baumgartner's[160] review led them to conclude that enclosing a duplicate questionnaire with a follow-up contact did not increase response rates beyond the effect of the reminder itself; across 32 studies with two contacts, the response rate was 62% in the 16 that included a duplicate questionnaire and 65% in the remainder.

The findings from the identified studies were equivocal.

In their complex experiment, Peterson and colleagues[184] manipulated the number (zero, one, two) and type (postcard only, letter plus duplicate questionnaire) of reminders. Average response rates across all strategies involving at least one duplicate questionnaire were 20%, significantly higher than the average (16%) for strategies involving only postcard reminders.

Swan and colleagues[197] examined the effect of including a duplicate questionnaire with both first (2 weeks after initial mailing) and second (2 weeks after first reminder) follow-up letters. For the first reminder, response rates were almost identical with and without inclusion of a duplicate questionnaire. However, for the second follow-up, the inclusion of another questionnaire led to a (just) significant increase in response rates *vis-à-vis* the group who had received a letter only.

As already noted, Roberts and colleagues[201] found no significant difference in response rates to a postcard first reminder and a letter with duplicate questionnaire.

Postal rates and types

Findings from previous reviews[158–162,168,202,203] regarding the impact of postage rates (e.g. first class versus second class) and type (e.g. hand-stamped versus franked or reply-paid envelopes) are equivocal and often contradictory. However, Dillman[1] was highly prescriptive regarding postage rates, advocating first class franked postage for outgoing mail (to enhance the image of "importance" and to facilitate forwarding to a recipient's new address or returning to sender) and addressed (rather than preprinted) first class business reply envelopes.

Identified studies

Ten studies were identified[191,192,196,204–211] in which postal rates and types were experimentally manipulated (*Table 19*; see p. 145). All were randomised

controlled trials but only three were on health-related topics.[208,209,211]

Comparisons involved:

- class of postage for both outgoing and return envelopes (one study)[211]
- outgoing mail (two studies), focusing on:
  - first class versus third class postage[207]
  - first class stamped envelopes versus bulk rate permit envelopes[196]
- return mail (seven studies) focusing on:
  - stamped versus business reply envelopes[191,192,204,205,208,209]
  - first class versus second class postage[210]
  - commemorative versus regular stamps.[205,206]

Outgoing and return mail

In a health survey, Cartwright and Windsor[211] compared the effect of first class and second class postage on both outgoing mail and return envelopes. The class of mail of outgoing and return envelopes was matched in all cases. There was no significant difference to final response rates. However, first class mailing yielded a faster response (percentage of questionnaires returned within 7 days: 27% for first class post; 13% for second class).

Outgoing mail

Findings from the two identified studies were equivocal.

Hopkins and Podolak[207] compared the impact of first class and third class mail in two separate experiments. In their first study, in which they also experimented with including a $1 incentive, a significant effect of postage rate was reported (however, the 95% CI for the RR includes unity). There was an interaction between postage rate and incentives. When an incentive was used, first class mail led to a significant improvement in response rates (68% versus 40%; RR = 3.26; 95% CI, 1.40 to 7.57). In the absence of an incentive, the difference in response rates was not statistically significant and the observed response rate was actually lower for first class mail (35% versus 39%; RR = 0.87; 95% CI, 0.54 to 1.40)). In their second study, in which no incentives were used, response rates were lower and there was no difference between first class and third class mail.

Kahle and Sales[196] compared the effect of using a first class stamp with that of using a bulk rate permit number on outgoing envelopes; the degree of personalisation was also manipulated. No overall effect of postage rate was found; nor, on controlling for whether the recipient's address was individually typed, did response rates for the stamped and bulk rate mailings differ significantly from each other.

Return mail

In all eight experiments identified by Linsky,[158] in which stamped and business reply return envelopes were compared, response rates were always higher for stamped envelopes. Linsky suggested that there may be a psychological barrier to throwing away an unused stamp because of the monetary value. However, business reply envelopes may offer a cost advantage because the postage cost is incurred only if the envelope is returned.

Findings from two of the five identified studies that compared stamped and business reply envelopes supported Linsky's conclusions; one yielded mixed findings and two found no effect of return postage type on response rates. This mix of findings held true for both health-related surveys and those on other topics.

Harris and Guffey[204] found that final response rates were (just) significantly higher for stamped envelopes, but replies were received more speedily with business reply envelopes (responses within 2 weeks of mailing: 92% for stamped envelopes; 98% for business reply envelopes). Harris and Guffey concluded that stamps appeared to be more cost-effective than permits, except for large surveys with relatively low anticipated response rates.

Jones and Linda[205] compared commemorative and regular stamps and business reply envelopes in a study in which sponsorship and nature of the appeal were also varied. Response rates were lower when a business reply envelope was used. However, the business reply envelope also yielded the lowest cost per returned questionnaire (49 cents compared with 64 and 68 cents respectively for regular and commemorative stamps). There were no significant differences in item non-response rates, or any evidence of response bias with respect to type of postage.

In a health-related survey, Corcoran[208] observed a (just) significant difference in initial response rates in favour of stamped over reply-paid envelopes. However, after a single postcard reminder, final response rates in the two groups were not significantly different. The cost per return was higher (55 cents versus 49.8 cents) in the group who received envelopes that were stamped.

Elkind and colleagues,[209] in a survey of professional psychologists, found no significant difference in

response rates between stamped and reply-paid envelopes, a finding echoed by Martin and colleagues.[191,192] In this latter study, the finding of no significant difference between stamped and business reply envelopes persisted on controlling for prenotification, follow-up or personalisation of cover letter; the additional cost of stamped envelopes was $0.19 per questionnaire sent.

Labrecque[206] looked at the effect of using commemorative stamps on the return envelope. Although a slightly higher response rate was observed for the commemorative stamp group, the difference did not reach statistical significance.

Harvey,[210] in a comparison of first and second class stamps, also found no effect of postage rates; response rates with first class stamped envelopes were in fact slightly lower.

Finally, the study by Jones and Linda[205] comparing commemorative and regular stamps did not demonstrate any significant advantage of the former over the latter.

Anonymity/confidentiality

The terms "anonymity" and "confidentiality" are often used synonymously but, as Zeinio[212] pointed out, they are not equivalent. Under conditions of anonymity, no individual identification appears on the questionnaire or interview schedule; it is not possible to link individual responses to a specific named person. Under conditions of confidentiality, an identification code (usually a number) appears on the questionnaire and those responsible for data collection can link this code to a named individual, but individual responses cannot be attributed to a specific person by anyone without access to the link between code and name; ethical principles also require that respondents' answers are not revealed to a third party without their explicit permission and that results are presented in such a way that individual respondents cannot be identified.

A fully anonymous approach precludes targeting initial non-respondents with follow-up contacts; any reminders have to be sent to the whole group, which clearly increases the cost of follow-up and runs the risk of antagonising or alarming those who have already responded. For this reason, some experts[180] suggest a combination of an unmarked questionnaire and an identifiable postcard to be posted back separately to indicate that the questionnaire has been returned.

Identified studies

Five studies[185,213–216] in which aspects of anonymity or confidentiality were manipulated in experiments were identified (*Table 20*; see p. 149). All were randomised trials but only one was on a health-related topic.[215] As noted above, "anonymity" and "confidentiality" were operationalised in quite different ways across these studies. For example, anonymity was variously portrayed as respondents not needing to sign their name on the questionnaire or in terms of identification by a code number only.

Findings from one study provided evidence in support of a positive effect of anonymity on response rates; one (which was health related) produced equivocal results, with the possibility of confounding by the impact of follow-up. The remaining three studies demonstrated no significant differences in response rates.

McKee[216] investigated the use of coded questionnaires with a covering letter stressing that the code number was "only so that we can follow-up people who don't respond" (in this respect, his experiment could also be viewed as a manipulation of the content of the covering letter). A single reminder was sent to all those who had received an uncoded questionnaire and to non-respondents who had received coded questionnaires. Both initial and final response rates were significantly higher in the group receiving coded, identifiable questionnaires. The hypothesis that topic involvement (the extent to which people are involved in the activities that are the subject of the survey) among identifiable respondents would be lower than among anonymous respondents (since the negative incentive of a threatened follow-up may motivate less interested individuals to respond) was also supported. Response quality, measured in terms of the percentage of closed questions answered, the total number of words used in answering open-ended questions, and the number of written comments made, did not differ significantly between the two groups. However, mean scores for the anonymous respondents on a 5-item scale of attitude towards the perceived importance of the organisation concerned were significantly higher. McKee[216] concluded that anonymity is likely to lead to sample composition bias and response bias in favour of respondents with a greater interest or involvement in the survey topic.

Campbell and Waters[215] hypothesised that complete anonymity would lead to higher response rates in a survey on a sensitive topic, namely the public's level of knowledge of AIDS. In six separate replications

they compared a numbered questionnaire with an unnumbered questionnaire accompanied by an assurance of total anonymity, indicating that the recipient was identifiable from the number and would be sent a reminder. There were no significant differences in response rates to the initial mailing between the unidentifiable and numbered questionnaires. Three weeks after the initial mailing, a reminder letter and duplicate questionnaire were sent to all non-respondents in the group that had received the numbered questionnaires; 43% subsequently returned a completed questionnaire, bringing the final response rate to 72% in the numbered questionnaire group, which is significantly different from the response rate of 49% with no reminders in the anonymous questionnaire group. Campbell and Waters[215] reported that, in a subsequent survey of health professionals, they used a numbered questionnaire but did not explicitly state in the covering letter that the numbering was to facilitate follow-up. They found that some recipients of reminders believed that they had been misled concerning the nature of the confidentiality; as a result, they recommended that the initial covering letter should be explicit about the purpose of numbering the questionnaires.

Jones[213] hypothesised that anonymity and sponsorship may interact with population characteristics in influencing response rates. In a very large study, each experimental "block" was a county; a wide range of socio-economic and cultural characteristics were known to exist across these counties. "Anonymity assurance" took the form of a statement in the covering letter: "Please note that you do not need to sign your name to this questionnaire. Your answers will remain completely anonymous." Overall response rates were very low (21%) and anonymity did not affect these. However, as hypothesised, anonymity assurance increased response rates for higher income populations, while the generally low response rate among populations in flux was further depressed by such an assurance. Contrary to expectations, there was no significant decrease in response rate under anonymity assurance among larger populations.

McDaniel and Rao[214] hypothesised that asking respondents to identify themselves by signing the completed questionnaire would have no effect on overall response rates (in comparison with an assurance of total anonymity), but that response quality (measured in terms of item omission, response error and completeness of answer) would be higher amongst those respondents required to identify themselves. Their rationale was that respondents would be more conscientious when they realised that they could be identified. However, their analysis showed that the only significant difference between the anonymous and identifiable groups was with respect to response error (the accuracy of responses to items that were verifiable by the researchers from documentary sources). Those respondents who were asked to sign their completed questionnaire gave more accurate answers (suggesting that they had taken more care in following the instructions to consult documentary sources).

Childers and Skinner,[185] in a covering letter, told one-half of their sample that their name and address (which was either preprinted on the return envelope or which they were asked to write on the envelope) would be used for research purposes only, while the other half were told only of the purpose of the survey. No significant differences in response rates were found between the two groups; nor were there any differences in response speed or item completeness.

Personalisation

Dillman[1] recommended personalisation of covering letters to "show regard for the respondent". His recommendations included individually addressed letters and envelopes, and hand-signed letters; he did not appear to consider that this may be infeasible or not cost-effective in large surveys.

However, findings from previous reviews on the impact of personalisation are equivocal. Linsky[158] identified 16 studies in which personalisation was examined; nine reported higher response rates for a personalised approach, four showed no differences, and three reported higher response rates for a non-personalised letter. Wiseman,[217] citing earlier work by Andreasen,[218] recognised a fundamental tension between personalisation and anonymity (although, in common with many, he confuses anonymity and confidentiality). He argued that putting a respondent's personal details on a questionnaire represents "deprivation of confidentiality" rather than "personalisation" of the questionnaire and suggested that this may explain the negative effect of personalisation found by Houston and Jefferson.[219]

Trice[220] also highlighted the possibility that personalisation may give rise to doubts about "confidentiality" and suggested that it "may not be a unitary function" but rather may consist of three distinct elements: salutation, body of letter and signature.

Identified studies

Eleven studies[185,191,192,196,206,221–227] were identified in which various aspects of personalisation were manipulated in an experimental design (*Table 21*; see p. 151). In all 11, the study design was a randomised controlled trial; however, only two involved surveys on health-related topics.[221,226]

Personalisation was operationalised in different ways across these studies:

- personalised versus form letter (six studies)[191,192,206,221,223,224,227]
- salutation and signature styles (one study)[225]
- addressing of envelope (four studies)[185,196,206,226]
- personalisation of appeal (one study).[222]

Only two studies reported any significant positive effects of personalisation, one in respect of personalisation of the covering letter and one in respect of personalisation of outgoing envelopes.

Personalised versus form letter

Green and Kvidahl[227] compared an individually printed covering letter bearing a personal salutation and address and a hand-signed signature with a mimeographed letter bearing the salutation "Dear Educator" and a facsimile signature. The response rate to the personalised letter was 9% higher, a statistically significant difference. With personalisation, response was also more rapid; more replies were received to the initial mailing and fewer to the final (second) follow-up. The cost per returned questionnaire was $1.603 for the personalised letter and $1.605 for the form letter.

The other five studies that manipulated personalisation of the covering letter found no significant differences between personalised and non-personalised letters.

In a health-related survey, on controlling for presence/absence of a social appeal and a deadline, Roberts and colleagues[221] found no effect on overall response rates of personalisation (addressing the letter to "Dear Dr (name)" or an open address). However, the highest initial response rate was for a personalised letter, specifying a deadline but with no social appeal; the highest final response rate was for a personalised letter with a deadline and a social appeal.

Labrecque[206] defined personalisation in terms of a hand-addressed outgoing envelope and a covering letter with a handwritten salutation and signature, and found no significant effects on overall response rates and no significant interactions between personalisation, status of the sender, and type of stamps used on outgoing mail.

Martin and colleagues[191,192] manipulated personalisation of the covering letter, as well as prenotification, follow-up and type of postage on return envelopes. Personalisation alone did not have a significant effect on response rates, but the interaction between personalisation and prenotification just reached statistical significance; personalisation increased the response rate among those who were prenotified.

Woodward and McKelvie[223] compared four different forms of address in an experiment that also involved manipulating the perceived "interest level" of the topic to the student recipients. Although the most familiar form of address (shortened version of forename plus surname) resulted in the highest response rate, the overall differences between the four groups were not statistically significant (and response rates in all four groups were low). In individual comparisons, the nickname/surname combination was significantly better than a box number alone.

Worthen and Valcarce[224] compared the impact of a personalised letter (individually typed, addressed to the recipient by name and personally signed by hand) and a form letter (mimeographed, addressed to "Dear Teacher" and with a facsimile signature). Although overall response rates were low, due to the timing of the survey, there was no significant difference in response rates to the initial mailing. Non-respondents to the first mailing were sent either a personalised or a form letter reminder 6 weeks later. No significant differences in response rates with type of reminder letter were observed; this was true of the entire sample and on controlling for the type of initial letter.

Salutation and signature styles

In the single study of salutation and signature styles detected, Green and Stager[225] compared the effects of a personalised and general ("Dear Educator") salutation and of a hand-signed or duplicated signature (considered to be less personal), using a factorial study design. Their analysis showed no main effects of salutation or signature. However, the interaction between salutation and signature approached statistical significance.

Addressing of envelopes

Kahle and Sales[196] manipulated personalisation of the outgoing envelope and the type of postage. In all cases, the recipient's name was individually typed on the covering letter. When addresses were individually typed, there was no significant

difference in response rates between first class and bulk rate outgoing postage; both yielded higher response rates than the combination of a preprinted address label and bulk rate mailing (although the contrasts were reported to be statistically significant, the 95% CIs for the RR include unity).

The other three studies focusing on how the address was added to the outgoing envelope found no significant effect of method.

As already noted, Labrecque[206] found no significant effects on overall response rates in his comparison of a hand-addressed outgoing envelope and a covering letter with a handwritten salutation and signature.

Childers and Skinner[185] experimented with personalisation of outgoing and return envelopes, as well as with the content of the covering letter. In a factorial design, half the sample received questionnaires in envelopes on which their addresses had been printed directly by computer, to simulate the effect of a typed address, and the remainder received envelopes with the address added on a computer printed adhesive label; it was hypothesised that the more personalised appearance of the computer printed envelopes would lead to higher response rates. For half of each group, the return envelope was personalised by computer printing of the respondent's name and address, while for the remainder a label was affixed asking the respondents themselves to add these details (supply of the sender's address is usual practice in the USA); it was hypothesised that the extra "cost" to the respondents of having to add their own details may lead to lower response rates. The findings showed no significant differences in response rates with level of personalisation of outgoing or return envelopes and no significant interaction effects. Nor did personalisation appear to lead to response bias, as measured by scores on 12 Likert scale attitude questions. However, response completeness was significantly associated with the personalisation of outgoing envelopes, with a lower rate of item non-response being observed for respondents receiving computer printed envelopes.

Wunder and Wynn,[226] in a survey of satisfaction among members of a health maintenance organisation, compared hand-addressed envelopes with those bearing a computer-generated address label. No significant differences in response rates or in time taken to respond were found. Nor did personalisation increase the quality of response measured in terms of item non-response rates and fullness of answers to an open-ended question.

Personalisation of appeal

Finally, Childers and colleagues[222] conducted a study in which the nature of the appeal to respond and the format in which it was presented were varied. In all cases the appeal was contained in a postscript to the covering letter. For half of the potential participants they used an offset-printed facsimile of a handwritten postscript, while in the remainder the postscript was typed. In two separate samples (of academics and of business practitioners) no significant difference in response rates was found between the two versions. Nor was there a significant interaction between the format of the postscript and the nature of the appeal. Furthermore, the format did not lead to significant differences in response completeness or in response bias (defined in terms of similarity of response patterns across treatments).

Covering letters

Zeinio[212] highlighted the role of a covering letter in persuading recipients to participate in a survey and recommended that the letter should anticipate and counter all arguments that a recipient may present against participation. In particular, he suggested that the letter needs to convey the "salience" or importance of the survey through the use of appropriate appeals and through the identity and prestige of the sender, and should stress the importance of a response from the targeted individual.

Sudman[180] suggested that, in surveys of professionals, the covering letter could usefully be accompanied by more extensive explanatory material, although he did not present any evidence to show whether the amount and nature of supporting documentation significantly affects response rates.

Dillman[1] advised that covering letters for household surveys should: explain what the study is about, emphasising its social usefulness; highlight why the sampled individual is important (and, if necessary, indicate which household member should answer the questions); provide an assurance of confidentiality, including an explanation of why identification numbers are being used; indicate how the results of the study will be used; offer a summary of the results of the survey; provide information on what the recipient should do if questions arise; and thank the recipient for their assistance. He also recommended that the title or job position of the sender should be included.

Identified studies

Fifteen studies were identified that met the inclusion criteria and in which aspects of the covering letter were experimentally manipulated (*Table 22*; see p. 155).[48,110,195,205,206,211,221,222,228–234] All were randomised controlled trials and five were on health-related topics.[211,221,228,229,233]

Comparisons involved:

- style of letter (one study)[230]
- characteristics of the signatory (three studies),[206,231,233] two of which also involved manipulation of the style of the signature, considered by Dillman[1] to be an aspect of personalisation
- nature of the appeal made in the letter (ten studies)[110,205,211,221,222,228,229,232–234]
- provision of time cues (one study)[48]
- specification of deadlines (two studies),[195,221] one of which also looked at the nature of the appeal.

Style of letter

Wagner and O'Toole[230] investigated the impact of non-traditional communication in an invitation to academic psychologists, asking about willingness to administer a survey to their students. Half of the sample received a traditional, personalised covering letter and a form on which to indicate their willingness to participate in the survey proper; the remainder received a humorous form, offering the incentive of a "free lunch" on return of the participation form. The traditional approach was significantly more effective than the humorous one in respect of the rate of return of the forms.

Characteristics of the signatory

Only one of the three studies identified showed a significant effect of the characteristics of the signatory of the covering letter on response rates.

Labrecque[206] experimented with the status of the sender (the owner versus the service manager of a marina) as well as personalisation and return postage rates. A (just) significantly higher response rate was observed when the signatory of the covering letter was the owner of the marina (a more prestigious position).

In a factorial design, Dodd and Markwiese[231] compared the effects of the status of the sender, sex of the sender, and personalisation of the signature. No significant effects of the demographic characteristics of the sender on response rates or questionnaire completion rates were found, and there were no significant interaction effects.

Also using a factorial design, this time in a health-related survey, Dodd and colleagues[233] examined the effects of single versus multiple signatories to a covering letter, as well as the colour of ink used for the signatures and whether a postscript was included in the letter. The number of signatories did not have a significant effect on response rates.

Style of signature

Dillman[1] strongly advocated that covering letters should be individually signed, using a blue ball-point pen, to avoid any impression of facsimile signatures. However, the findings from the two identified studies are equivocal.

Dodd and Markwiese[231] compared the impact of a handwritten and a facsimile signature; the accompanying letters were all photocopied and were not personalised in any other way. Although overall response rates did not vary significantly with type of signature, those receiving a hand-signed covering letter were more likely to return a completed (as opposed to blank) questionnaire.

Dodd and colleagues[233] experimented with the colour of ink in which the signatures on the covering letter were written, in a factorial design study in which the number of signatories and the presence/absence of a postscript appeal were also manipulated. Ink colour (bright green or regular blue) had no significant effect on response rates.

Nature of appeal

Theories of individual motivation have underpinned attempts to increase response rates by manipulation of the nature of the appeal made in the covering letter. For example, McKillip and Lockhart[229] drew on Katz's functional theory.[235] According to Katz, there are four motivational bases for an individual respondent's attitude towards a questionnaire topic: utility, reflecting past experience of rewards and punishments; value-expression, reflecting reference groups and other symbols that give positive expression to individuals' self-image; knowledge, reflecting individuals' desire to make sense of their world; and ego defence, reflecting individuals' desire to avoid confronting painful stimuli. Drawing on this theoretical framework, McKillip and Lockhart[229] argued that the nature of the appeal presented in a covering letter should make a convincing link between the questionnaire topic and an important motivational concern of the respondent.

"Reactance theory" provides another theoretical perspective of relevance to covering letter appeal.[236–238] A covering letter with an overt appeal

to respond may be perceived as a threat to the recipient's freedom of choice regarding response; by deciding not to respond, the threatened freedom is restored. In contrast, stressing the recipient's personal freedom in making the decision regarding response may reduce such reactance effects.

Social utility appeals

Five of the identified studies included an examination of the impact of a "social utility" appeal, as advocated by Dillman.[1] Findings from these studies were mixed and indicated no consistent advantage in favour of such an appeal.

In a health-related survey, Roberts and colleagues[221] examined the impact of a social utility appeal (compared with no appeal), emphasising the relevance of the research to fellow dental practitioners. No main effects on initial or final response rates of including a social utility appeal were found in this study, in which deadline specification was also manipulated.

McKillip and Lockhart[229] conducted two studies among student populations in which they drew on Katz's functional theory.[235] In the first study they compared the effects of what they termed "utility", "value-expression" and "knowledge" appeals: the utility appeal emphasised the value of the study to the respondent as an individual; the value-expression appeal emphasised its value to students and the university in general; and the knowledge appeal focused on the contribution to general and personal knowledge bases. Although the utility appeal evoked the best response and the value-expression appeal resulted in the poorest response among undergraduates, the differences did not reach statistical significance (contrary to the conclusions of McKillip and Lockhart themselves[229]). Among postgraduates, only the utility and knowledge appeals were used and there was no significant difference in response rates between them. In the second study, the same researchers tested whether a combined knowledge–utility appeal would increase response rates over a utility appeal alone. Among undergraduates, a higher response rate was obtained with the utility appeal alone, but the opposite was found in the postgraduate sample; however, in neither sample did the differences reach statistical significance (again contrary to what was stated by McKillip and Lockhart[229]). It should also be noted that the terminology used by McKillip and Lockhart is at odds with that used by other researchers; their "utility" appeal is what others term "self-interest" or "egoistic", while their "value-expression" appeal is similar to a "social utility" or "altruistic" appeal.

Jones and Linda,[205] in a factorial design study, compared the impact of three levels of appeal: an "altruistic" or "social utility" appeal, emphasising the value of a response to science or "society"; a "self-interest" appeal, stressing the benefit to the respondent as an individual; and an appeal highlighting the value of the survey to the sponsor. In this study, the identity of the sponsor and the type of return postage were also manipulated. No difference in response rates between the social utility and self-interest appeals was observed, but both were higher than the rate for the sponsor-interest appeal (although the 95% CIs for the RR include unity). No significant interactions with the other experimental variables were observed. However, significant differences in item non-response rates across the three groups were observed, with the highest rate of item omission in the sponsor-interest group and the lowest in the altruistic appeal group, but there was no evidence of significant response bias (measured in terms of answers to 37 questions) with respect to type of appeal.

Childers and colleagues[222] compared egoistic, help the sponsor and social utility appeals in two separate samples, one of academics and one of business practitioners. In all cases the appeal was given prominence by placing it in a postscript to the covering letter. In both samples a control group received a covering letter with no such postscript. In the academic sample a significant difference was observed in comparing the three appeals; further analysis showed that, although there was no significant difference in response rates between the egoistic and help the sponsor appeals, both led to a significantly higher response that the social utility approach. However, the highest response rate of all was in the control group, to whom no appeal was made, suggesting that explicit appeals may act as a disincentive, at least among academic populations. In the sample of business practitioners, the type of appeal had no significant impact on response rates. In neither sample was the type of appeal significantly related to response completeness or response bias.

Jones and Lang,[110] in a survey of house purchasers, hypothesised that an egoistic appeal would induce a higher response rate than a social utility appeal and would reduce sample composition bias, but that the social utility appeal would decrease response bias. They recognised the possibility of interactions between the nature of the appeal and other variables manipulated (sponsorship, the number and timing of contacts with sampled individuals, and the order in which two sets of questions were presented), and with socio-

economic status. However, they found no significant difference in response rates between the social utility and egoistic appeal. Other results were also contrary to expectations: the egoistic appeal yielded a sample biased in favour of purchasers of higher-priced houses; the social utility appeal yielded a sample intermediate between the egoistic appeal and the underlying population, and not significantly different from either; and neither type of appeal led to significant sample composition bias with respect to date of purchase. Finally, although the nature of the appeal alone did not lead to significant response bias (i.e. mismatch between reported and actual house price or purchase date), Jones and Lang[110] found a significant interaction between sponsorship and type of appeal. A combination of private agency sponsorship and an egoistic message led to "telescoping" (i.e. a bias in favour of reporting the purchase date as more recent than it really was).

Reactance-inducing appeals
Two studies examined whether reactance theory could provide an explanation for response behaviour in experiments in which incentives were also manipulated; the findings were mixed.

Using a factorial design in which an enclosed incentive of $1 was also manipulated, Biner[232] compared a covering letter emphasising the importance of the survey and how essential a response was with a version stressing that whether or not the recipient responded was a matter of personal choice. The "essential response" version was expected to induce reactance and thus result in lower response rates. Overall response rates in the group receiving the reactance-inducing version were significantly lower (although the upper bound of the 95% CI for RR is unity). However, analysis of the interaction between type of appeal and provision of an incentive showed that the observed effect of appeal type was mainly attributable to a large difference in the group receiving a $1 incentive.

In a subsequent study, Biner and Barton[234] manipulated both the magnitude of the incentive (25 cents versus $1.00) and the nature of the appeal, using a factorial design. One version of their covering letter stated that the enclosed money was to induce an obligation to respond, while the other portrayed the incentive in the traditional "token of appreciation" manner. Overall response rates for the "obligatory" letter were higher; however, further analysis showed that this effect was due almost entirely to a large difference among those receiving a $1.00 incentive. These results seemed to support an equity interpretation (i.e. that recipients sought to restore a feeling of equity between themselves and the researchers by returning their questionnaires to balance the financial "compensation"). However, Biner and Barton[234] speculated that reactance might still have occurred, leading individuals in the "$1.00 obligatory" group to return their questionnaire, but with bogus answers. A comparison of response patterns across the four groups (defined by size of incentive and type of appeal), however, showed no significant differences, leading to the conclusion that the reactance interpretation was not applicable.

Miscellaneous aspects of appeals
The remaining three studies investigated miscellaneous aspects of appeals, with mixed findings.

Salomone and Miller[228] experimented with four different appeals: an appeal to the professionalism of the potential participants; a humorous appeal; an emphasis on the importance of the individual respondent; and a presentation of token compensation (enclosure of 25 cents). They observed differences in response rates both to the initial mailing and subsequent two follow-ups. However, these differences were all attributable to the offer of token compensation; there were no differences in response rates between the other three types of appeal.

Dodd and co-workers,[233] using a factorial design study, examined the impact of a simple appeal in the form of a handwritten postscript saying "Please help!". Including the postscript did not significantly affect response rates.

In a postal survey of the general public, with the aim of screening patients who had attended hospital outpatient departments or who had consulted a GP but had not been referred, Cartwright and Windsor[211] experimented with the inclusion or exclusion of a question asking if the person would be willing to assist the researchers again in the future, and requesting the provision of a telephone number from those willing to help in further research. Response rates were significantly lower when this question was included.

Time cues
Hornik[48] postulated that response rates and response quality would be higher among respondents receiving a time cue indicating that the time required to complete the questionnaire would be short. His study involved a 39-item postal questionnaire, estimated to take 28 minutes to

complete. Three different covering letters were used: in the first, recipients were led to believe that the questionnaire would take 20 minutes to complete; in the second, they were told that it would take 40 minutes to complete; in the third, no time cue was provided. Significant differences in response rates between the short and long time cues, and between the short time cue and no time cue were found, which were in favour of the short time cue. Answers to an open-ended question indicated that respondents' perceptions of the time they had spent in completing the question-naire were lowest for the group provided with a short time cue. Response speed (i.e. days to respond) was also significantly higher among respondents who were provided with a short time cue. However, no significant differences in response quality, as indicated by the level of item non-response, were found. Hornik[48] also tested for response bias, hypothesising that the short time cue may encourage respondents to rush through the later questions; no differences in mean item scores were found between the three groups, indicating that response distortion had not occurred. He recognised that the manipulation of time cues may involve the provision of inaccurate information, giving rise to ethical concerns.

Specification of deadline

Findings from the two identified studies provide no conclusive evidence in favour of or against specification of a deadline.

Nevin and Ford[195] tested four different versions of covering letter: one specified no deadline while the other three specified deadlines of 5 days, 7 days and 9 days respectively. There was a significant linear trend in response rates across the three groups in which a deadline was specified, suggesting that a longer deadline "has a favourable influence on overall response rates". However, on comparing individual deadlines to the control of no deadline, the RRs all included unity. Comparisons of the three deadline groups indicated that the only significant difference in response rates was between the 5-day and 9-day deadlines. Response rates for the 7-day deadline and no deadline were very similar, suggesting that respondents may implicitly assume a deadline of 1 week in the absence of any cue to the contrary. However, contrary to expectations, specifying an explicit deadline did not lead to a more immediate response; response rates at 5 days after posting were very similar across all four versions of the covering letter. The different version of covering letter did not have any significant impact on item non-response rates or on response bias measured in terms of how the questions were answered.

Roberts and colleagues[221] found that specifying a deadline (through a statement in the covering letter saying "If we have not heard from you in 3 weeks, we will contact you again.") led to a significantly higher response rate to the initial mailing and subsequent to a first mail follow-up (after 4 weeks), but that there was no significant difference in final response rates (after a second follow-up at 8 weeks). They concluded, however, that the speedier response evoked by the specification of a deadline was cost-effective, reducing the cost of initial follow-up by approximately 25%.

Sponsorship

Heberlein and Baumgartner's[160] review found that government-sponsored surveys elicited higher response rates; their simple regression model showed that, on controlling for number of contacts and saliency, government-sponsored surveys should yield an additional 12.4% of responses.

Identified studies

Five studies were identified[110,150,205,213,239] in which aspects of sponsorship were experimentally manipulated (*Table 23*; see p. 161). Four out of the five were randomised controlled trials; the fifth was a non-random concurrent controlled study. Two were on health-related topics.[150,239]

Findings from three of these studies showed overall positive effects of sponsorship, while results from the other two were equivocal.

Jones and Linda[205] manipulated the letterhead on the covering letter and the address on the return envelope to examine the effect of three sponsorship conditions: a government agency, a university department, and a (fictitious) market research company. Response rates across the three groups were significantly different, with the best response rate obtained for university sponsorship and the poorest for the market research company. There was no significant interaction with the other factors manipulated in this study, namely the type of appeal made in the covering letter or the type of return postage. Nor were there any significant differences in item non-response rates, or any evidence of response bias with respect to sponsorship.

Jacoby[150] examined differences in the rate and speed of response when questionnaires seeking users' views of GP services were sent out by an independent research unit or by the local FPC, which at the time had responsibility for the organisation and administration of GP services.

Questionnaires were returned to the institution that sent them. In each of the two areas where the experiment was carried out, the FPC achieved a significantly higher response rate. Overall, the speed of response (operationalised as the percentage of returns received within 3 weeks of initial mailing) was also faster for the FPC; this was due to a large and highly significant difference (58% versus 37%) in one of the areas. However, levels of expressed satisfaction with GP services did not vary with the identity of the institution sending the questionnaires, suggesting that the identity of the sender did not lead to response bias.

Smith and colleagues[239] compared the impact on response rate of the recipient's own GP sending the introductory letter to the survey with that of the letter coming directly from the research unit conducting the survey. Both crude response rates (inclusive of questionnaires that could not be delivered by the Royal Mail) and adjusted response rates, after a single reminder, were significantly higher when the letter was on headed paper from the GP, although the CIs for initial response rate included unity; the greatest difference was among male respondents aged 40–49 years. Smith and colleagues[239] concluded that the enhanced response rates were probably due to the implicit professional relationship between GPs and patients; patients may consider that they know the GP better than they do the researcher, or they may feel some obligation towards the doctor.

Jones[213] tested whether the impact of sponsorship (university versus government) and anonymity on response rates was mediated by population characteristics. Although a significant effect of sponsorship on overall response rates was observed (magnitude and direction not reported), only one of his hypotheses regarding interactions was supported. The university sponsor experienced higher response rates in the area immediately surrounding that university in comparison with response rates in a competing university's area. Jones concluded that the benefit of university sponsorship may be quite localised.

In a postal survey of house purchasers, Jones and Lang[110] hypothesised that university sponsorship would yield higher response rates than private agency sponsorship, and that it would reduce sample composition bias and decrease response bias. The response rates were higher under university sponsorship, but findings on sample composition bias were mixed. Univariate analysis indicated that private agency sponsorship led to over-representation of those whose houses had been more expensive, but to no significant response bias with respect to date of purchase. No bias with respect to house purchase dates was identified for either form of sponsorship. Multivariate analysis showed that, although the mean purchase price among respondents to university sponsorship was closer to the underlying population mean, the estimate from this sample did not differ significantly from the estimates derived from the more absolutely biased sample resulting from private agency sponsorship. Contrary to expectations, sponsorship alone did not lead to significant response bias (i.e. mismatch between reported and actual house price or purchase date). However, there was a significant sponsorship by appeal interaction; a combination of private agency sponsorship and egoistic message led to "telescoping" (i.e. reporting the purchase date as more recent than it really was).

Saliency/subject matter

In their review, Heberlein and Baumgartner[160] defined a "salient" topic as one "which dealt with important behaviour or interests that were also current". Measuring saliency on a 3-point scale, they found a significant positive correlation with response rate ($r = 0.427$, $p < 0.001$). In their simple regression model, taken together, saliency and number of contacts accounted for 50.5% of the overall variance in final response rates.

Identified studies

Three studies were identified in which saliency was manipulated experimentally (*Table 24*; see p. 163). Two were trials on non-health-related topics;[223,240] the other was on a health-related topic but used a non-random concurrent control.[178] Two showed a positive effect of saliency on response rates, while the third found no significant differences.

Dommeyer[240] compared an "interesting" questionnaire (the 44-item Mind Inventory Catalogue) and an "uninteresting" questionnaire (the 55-item Tax Survey) in an experiment with a group of business studies students. Response rates to the "interesting" questionnaire were significantly higher.

Hovland and colleagues[178] examined whether questionnaire content had an impact on response rates. In their study, 200 dentists received a 9-item questionnaire asking for their opinions on the value of and need for basic science education, while a further 200 dentists were sent a 20-item questionnaire on their knowledge of dental drug costs. Initial response rates to the attitudes

questionnaire were significantly higher. After aggressive follow-up (two postal reminders plus telephone follow-up), response rates increased to 98% for those receiving the attitude questionnaire, and 95% for the knowledge questionnaire. However, even with the relatively low response rates to the initial mailing, no evidence of non-response bias was found; there were no significant differences in mean attitude scores or mean knowledge levels between early and late responders.

Woodward and McKelvie[223] experimented with questionnaires judged to be of "high" and "low" interest to the recipients (students of business and social science). Contrary to their expectations, the "low"-interest questionnaire attracted a higher overall response rate. The difference in "interest rating" was greater for the business students, but even among this group there was no significant effect on response rates.

Incentives

Hansen[241] suggested that self-perception theory[188,242] could explain survey participants' response to incentives; if behaviour is perceived to be influenced by plausible external causal factors (such as incentives), the individual should reject internal motivation as the cause of the behaviour. This would mean that, although the offer of an incentive may enhance response rates, it could lower the quality of response; in other words, those stimulated to reply by an incentive rather than because of intrinsic interest in the topic may be less motivated to give thoughtful answers, thereby bringing about response bias.

Furse and Stewart[243,244] discussed the applicability of dissonance theory to decisions regarding participation in postal surveys. They suggested that an enclosed financial incentive could cause feelings of dissonance with respect to decisions regarding reading and completing the accompanying questionnaire; to throw away the money would seem wasteful, while to keep it without completing the questionnaire would seem unethical. They speculated that an enclosed non-monetary incentive of low worth would not create the same sense of dissonance and may therefore be less effective in enhancing response rates, while a promised incentive would be more likely to be perceived as compensation for task completion. However, a fundamental flaw in this argument has been highlighted by Biner and Barton.[234] They pointed out that dissonance is a post-decision phenomenon and therefore should not affect decision-making with regard to survey participation.

Biner and Barton[234] therefore suggested that survey response behaviour, particularly the effect of incentives, could be more readily explained by two other theories. Equity theory[245] postulates that, when an individual feels overcompensated for an action, feelings of guilt are aroused. To reduce these feelings, the individual seeks to restore equity. In the case of survey research, this could be achieved by responding to the request to participate, thus increasing response rates. This theoretical explanation assumes that the incentive will be perceived as conveying a sense of obligation rather than of coercion. If the enclosure of an incentive is seen to be coercive, the second theory – reactance theory – could come into play. This proposes that, when faced with a threat to behavioural freedom, individuals will experience a state of arousal ("reactance") and will be motivated to reduce this reactance by restoring the freedom that is under threat. If this was the case, an incentive may perversely lower response rates.

Berry and Kanouse[246] speculated that the size of the incentive payment could be seen as a cue to the importance of the survey. A large incentive may be taken to indicate a significant research budget, implying an important study. The size of the payment could also be perceived to indicate the value of a completed questionnaire to the researcher. Payment in advance, Berry and Kanouse[246] argued, may be perceived as a signal of the researcher's trust in the target respondent. It may also be seen to initiate an exchange transaction,[144] which respondents feel obliged to complete.

Heberlein and Baumgartner's[160] review suggested a linear trend for incentives but there were no significant zero-order correlations (however, most of the studies they included used no incentives). Armstrong,[247] in a review devoted to monetary incentives in postal surveys, concluded that enclosed monetary incentives lead to enhanced response rates (particularly when the enclosure is with the initial mailing) and that the greater the size of the incentive, the greater the increase in the response rate. Hopkins and Gullickson[248] also confined their review and meta-analysis to the effects of monetary incentives. In an examination of 62 studies, involving 85 comparisons, they showed that response rates increased by 19% on average when a monetary incentive was enclosed; when the incentive was promised, the average increase was 7%. Larger incentives had a greater impact. These trends were consistent regardless of salience of the survey topic or the nature of the study population (general versus professional). The impact of incentives remained significant even

in studies with follow-up mailings. However, the impact was attenuated by poor survey design and implementation, in particular when the covering letter did not present the incentive as a gratuity (rather than as compensation).

Identified studies

The impact of incentives on response rates has been extensively researched. A total of 22 studies were identified on this topic (*Table 25*; see p. 164).[157,163,179,182,207,228,232,234,241,244,246,249–259] All but two were randomised trials; eight were on health-related topics.[179,228,246,250,252-254,256] Several involved manipulation of multiple aspects of incentives (e.g. magnitude and timing of delivery); the subsections below reflect the principal foci of the identified articles.

The comparisons were of:

- some form of incentive versus no incentive at all (19 studies)[157,163,182,208,228,232,241,244,249–259] (note that, although a comparison of no incentive versus incentive was possible for all of these studies, it was not always the focus of analysis by the original researchers)
- financial versus non-monetary incentives (two studies)[182,241]
- enclosed versus promised incentives (seven studies)[182,244,246,250,251,257,258]
- size of incentive (five studies)[179,234,244,250,258]
- other aspects of nature of incentive (two studies)[163,259]
- the appropriateness of equity theory and reactance theory in explaining response behaviour (one study).[234]

The majority of the identified studies showed a positive effect of incentives.

Incentive versus no incentive

Thirteen of the 19 studies in which some form of an incentive was compared with a "no incentive" control group showed a positive effect of incentives on response rates; three of these were health related.

In a health-related survey, Salomone and Miller[228] examined the effect of enclosing a 25 cent coin with the initial mailing and drawing attention to this incentive in the covering letter. Initial and final response rates were significantly higher among the group receiving the incentive.

In a survey of respiratory illness in children, Woodward and colleagues[254] offered the opportunity to enter a draw to half of the recipients of the questionnaire, where the prize was a voucher for a restaurant meal to the value of A$100. After reminders, the response rate among those who had been offered the incentive was (just) significantly higher than that for the control group who had received no incentive.

Weltzien and colleagues,[256] in a survey of satisfaction amongst ex-clients of a mental health service, experimented with enclosing a token incentive of 2 cents. Although overall response rates in their survey were very low (21%), a significantly higher response was obtained from those to whom the incentive was provided.

In a survey of the drinking, smoking and dietary practices of women postpartum, Little and Davis[250] compared promised and enclosed incentives of varying magnitudes. The response rate for all incentive groups combined was 69%, significantly higher than the 59% obtained when no incentive was provided.

In a survey of industrial safety engineers, Hansen[241] compared a group receiving a small incentive (25 cents or a ball-point pen) to a no incentive control group. Response rates for each incentive group individually and for the two groups combined were significantly higher than for the group receiving no incentive.

Furse and Stewart[244] experimented with enclosed and promised incentives of varying magnitudes. Response rates across all the incentive groups ranged from 56% to 78%. The average response rate in the incentive group was 70%, significantly higher than the response rate of 54% in the no incentive control group.

Hopkins and Podolak[207] examined the impact of enclosing a $1 bill in a survey in which a low response rate was anticipated. They found a significant difference in overall response rates between those who did and did not receive the incentive. In this study, the type of mailing for outgoing questionnaires was also manipulated and a significant interaction between postage rate and incentive was found. The incentive had no effect on response rate when third class (bulk) mailing was used, but it had a large effect when used in conjunction with first class postage. Costs per returned questionnaire were: $2.06 for first class mail plus incentive; $1.13 for first class mail without incentive; $3.13 for third class mail plus incentive; and $0.67 for third class mail without incentive.

Blythe[255] experimented with offering respondents the opportunity to participate in a lottery if they

returned a completed questionnaire; prizes were nine $20.00 gift vouchers and a weekend break. Response rates at 3 weeks after initial mailing (after a single reminder) were significantly higher in those offered entry to the lottery. Two subsequent reminders, in which all non-respondents were offered the opportunity to participate in the lottery on return of a completed questionnaire, brought in an additional 40 respondents, yielding a final response rate of 66%.

Trice,[157] in a survey of satisfaction among hotel guests, experimented with an incentive of a $1.00 reduction in the hotel tariff for respondents. Although overall response rates were low (as is typical of such surveys of customer satisfaction), the provision of the incentive led to a significantly higher response. There were no significant interactions between the provision of the incentive and any of the other variables manipulated in this study (timing of survey, number of questions and provision of space for open comments).

Biner[232] used a factorial design to examine the effects of an enclosed $1 bill and two versions of a covering letter appeal and found that the provision of an incentive significantly increased response rates, both overall and on controlling for type of appeal. A significant interaction between provision of an incentive and type of appeal was also observed. Biner[232] speculated that the enclosure of the incentive might have induced recipients to pay closer attention to the letter, either by grabbing their attention or by inducing a feeling of obligation to read it.

Hubbard and Little,[258] in a survey on satisfaction with financial services, compared enclosed and promised incentives of varying sizes. The response rate for all the incentive groups combined was 53%, significantly higher than the 41% response rate achieved in a no incentive control group.

Brennan[259] reported on a series of five experiments in New Zealand involving the comparison of a 50 cents incentive enclosed with the first mailing with no incentive. The survey topics and the number of follow-ups varied from study to study. In all five replications a higher response rate was obtained for those to whom an incentive was given; the difference reached statistical significance for three of out of the five. Using an incentive also increased the speed of response and reduced the need for follow-up; response rates to the initial mailing were between 7% and 21% higher in the incentive group. Overall refusal rates (i.e. the percentage returning a blank questionnaire) ranged from 2.7% to 8.1% when no incentive was included and from 4.2% to 8.9% for the incentive group. The cost per incremental return ranged from $0.76 to $4.78.

Gajraj and colleagues,[182] in a survey of customers of a public utility company, compared type and timing of incentives, using a no incentive control group. The response rate for all incentive groups combined was 48% compared with 34% in the control group, a statistically significant difference.

Two of the six studies showing no significant effect of incentives were health related.

Cook and co-authors[252] examined, in a 5-item initial survey of drug education programme administrators, the impact of a promised $100 incentive for those agreeing to participate in further research. The offer of an incentive had no significant impact on response rates or on willingness to participate in more detailed research.

Mortagy and colleagues[253] evaluated the effect of a raffle with a total of £100 in prize money. Initial response rates (i.e. without a reminder) did not differ significantly between the two groups. After a single letter reminder, which included a duplicate questionnaire but did not mention the raffle, the final response rate was 73% for the incentive group and 72% for the control group, a non-significant difference.

Whitmore[249] found no difference in response rates between those provided with a small material incentive (a key ring) and those receiving no incentive. A particular focus of this study was the possibility of response bias. Out of 83 items, the response patterns for those provided with an incentive were significantly different (at the 0.05 level) from those not receiving an incentive for only one item, a finding that could have occurred by chance given the large number of comparisons.

In a survey of business people, Paolillo and Lorenzi[251] focused primarily on enclosed versus promised incentives, but included a control group receiving no incentive at all. On comparing all incentive groups combined with this control group, the difference in response rates (46% versus 36%) failed to reach statistical significance.

In a household survey, Dommeyer[257] also experimented with enclosed and promised incentives of varying sizes, and compared these with a no incentive control group. The average response

rate across all incentive groups was almost 38%, compared with 37% in the control group, a non-significant difference.

Suhre,[163] in a survey of school principals, reported that response rates were similar in groups with promised incentives and those to whom no incentive was offered.

Financial versus non-monetary incentives

Findings from the two studies that compared monetary and non-monetary incentives were mixed.

Hansen[241] experimented with a small financial incentive (25 cents), a non-monetary incentive of similar value (a ball-point pen) and no incentive. Drawing on the self-perception theory outlined above, he hypothesised that response rates would be higher among the incentive groups but that response quality would be higher in the group not offered an incentive. His findings bore out these hypotheses. Response rates were higher in the two incentive groups than in the no incentive control group; the monetary incentive was more powerful than the pen. However, response completeness to both open-ended and closed questions, and the quality of response to open-ended questions were both better in the no incentive group. There was no evidence of response bias measured in terms of the distribution of response to closed questions or the suggestions for product improvement elicited by open-ended questions. Response speed was significantly lower in the group who received a non-monetary incentive.

Gajraj and colleagues' design[182] allowed the comparison of 50 cents, a pen and a lottery entry, whether enclosed or promised. Significant differences were found between an enclosed pen and an enclosed 50 cents (41% versus 62%), and between an enclosed pen and an enclosed lottery entry (41% versus 55%).

Enclosed versus promised incentives

Results from the seven identified studies supported findings from earlier reviews[247,248] that enclosed incentives are more powerful than promised incentives in enhancing response rates.

Berry and Kanouse[246] compared prepayment and postpayment of a $20 incentive in a postal survey of physicians. The timing of the payment had a significant effect on response rates after reminders (78% for prepayment versus 66% for postpayment). A subgroup of the postpayment group who had received a telephone reminder and had promised to return their completed questionnaire were sent their cheque in advance of the return of the questionnaire; the final response rate for this subgroup was 77%, indicating that prepayment was effective even when used late in the contact process. Prepayment also reduced the need for follow-up; only 44% of those in the prepayment group required a reminder, compared with 62% in the postpayment group. No significant response bias with respect to a range of demographic variables was observed between the prepayment and postpayment groups but overall non-response bias was observed (physicians in solo practice or partnerships, and those who were not board certified were under-represented). Among the prepayment group, the vast majority (95%) of those who completed a questionnaire cashed their cheques, while 26% of non-respondents in this group did so. The average payment per completed questionnaire was $21.45 in the prepayment group and $19.92 in the postpayment group.

Furse and Stewart[244] examined whether offering a donation to a charity of the respondent's choice (to be selected from a supplied list) would increase response rate over the effect of a personal enclosed incentive and over a no incentive control. The response rate for all incentives combined was significantly higher than that for the control group. However, in the absence of any enclosed incentive, a promised donation did not result in a significantly higher rate of response than no incentive at all. Overall, these researchers concluded that there was no significant effect on response rates of the charitable incentive. Incentives did not lead to significant non-response bias (measured in terms of the demographic characteristics of respondents and non-respondents); nor did they have a signifi-cant impact on speed of response or on item non-response rates. Cost–benefit analysis showed that an enclosed personal incentive alone was more cost-effective than the use of a $1 charitable donation, either alone or in combination with a personal incentive.

Paolillo and Lorenzi[251] compared an enclosed incentive of $1 with a promised incentive of either a guaranteed $2 or entry in a lottery for prizes to the value of $50, $30 and $20 per 100 respondents; a control group, to whom no incentive was provided, was also included. Significant differences in response rates across the groups were found. The response rate for the group receiving the small enclosed incentive was significantly higher than that for any of the other groups; none of the other pair-wise comparisons reached statistical significance, although the promised incentive of $2 elicited a higher response than entry in a lottery.

Hubbard and Little[258] sought to expand the work of Furse and Stewart[244] described above. They hypothesised that: there would be no significant difference in response rates between no incentive at all and a promised donation of $1.00 to the charity of the respondent's choice; that enclosed personal incentives of $0.25 and $1.00 would significantly increase response over the no incentive condition; and that promised entry in a lottery for a prize of $200 would also result in a significantly higher response rate. All three hypotheses were supported by their findings. The response to the lottery incentive was not significantly different from that of a $0.25 enclosed incentive; the cost per actual response for the lottery entry was lower than that for the enclosed cash incentives, although the marginal cost (cost per incremental response over the no incentive condition) was higher. No biasing of results by incentive was observed; nor did the provision or type of incentive significantly affect response quality or response speed.

Dommeyer[257] examined different forms of a monetary incentive. Three of the six groups received an enclosed incentive of 25 cents in the form of a coin, cheque or money order; a fourth group was offered a promised "early bird" incentive, whereby $25 would be shared amongst all those responding within a week of the initial mailing; members of a fifth group were promised that their names would be entered in a lottery for a prize of $25 on receipt of a completed questionnaire; a sixth control group received no incentive. The author hypothesised that each of the monetary incentives would produce a higher response rate than that observed in the control group, and that those receiving the "early bird" incentive would respond more quickly. He also anticipated that respondents in the no incentive group would have higher scores on the Morality–Conscience–Guilt scale[260] because "guilt-sensitive" individuals would be likely to respond out of a feeling of moral obligation, while a monetary inducement may be sufficient to prompt a response from individuals who are not motivated by such moral scruples. Preliminary analysis of variance showed that there were some differences in response rates across the six groups; on decomposition, this was found to be due to a significant difference between the 25 cent coin group (response rate 50%) and the lottery group (response rate 30%). The hypothesis of higher response for any incentive compared with none was not supported. However, the three enclosed incentives combined (coin, cheque, money order) did lead to a higher response rate than for the two promised incentives ("early bird", lottery). The findings did not support the hypotheses of a more speedy response to the "early bird" incentive or of higher Morality–Conscience–Guilt scale scores in the no incentive group. The interventions did not differ significantly in terms of item omi-sion and there was no evidence of response bias. The most cost-effective approaches were no incentive and the 25 cent coin.

Little and Davis[250] examined the response of pregnant women to various amounts of enclosed or promised monetary incentives. Findings showed that the enclosure of cash was superior to the promise of cash, or to no incentive at all, in enhancing response rates. A promised incentive led to significantly higher response rates than no incentive, but the response rate to a $2 promised incentive was not significantly different to that for a $1 promised reward. Little and Davis[250] commented that payment for service prior to the service being rendered is illegal under some circumstances. The return of a signed consent form with the questionnaire was indicative that the "service" (in this case the questionnaire) had been rendered, but few of the non-respondents in the "incentive enclosed" group returned the $1 they had been sent. They also pointed out that the true cost of promising or enclosing money was greater than the "face" cost; in their study both "$1 promised" and "$1 enclosed" cost the researchers $1.34 (due to costs of follow-up of non-respondents, and of writing and posting cheques).

Gajraj and colleagues[182] experimented with different types of incentive (none, monetary, gift of a pen, entry in public lottery) and with enclosing or promising the incentive. In comparison with no incentive, all types of incentive, whether promised or enclosed, led to higher response rates. The differences were statistically significant for the enclosed monetary (50 cents) and enclosed lottery incentives. As hypothesised, enclosing an incentive led to a higher response rate than promising one on return of a completed questionnaire (although the differences for the pen and lottery incentives were not statistically significant). Response rates for both the monetary and lottery enclosed incentives were significantly higher than for the gift enclosed incentive, in line with expectations; however, the difference between response rates for the monetary and lottery enclosed incentives did not reach statistical significance. All incentives led to faster return of questionnaires, measured in terms of the mean number of days to reply; the difference reached statistical significance for all but the enclosed lottery incentive. Enclosing an

incentive did not significantly increase the speed of response over promising the equivalent incentive. A monetary incentive led to a faster response than a gift incentive, but a lottery incentive did not increase response speed by comparison with either a gift or money. The lowest cost per response was obtained for the 50 cents enclosed incentive.

Size of incentive

Findings from the five studies that compared the effect of different levels of incentive were equivocal.

In comparing enclosed incentives of 25 cents and $1, Hubbard and Little[258] found a significant difference in favour of the larger incentive, a finding echoed by Biner and Barton.[234]

Little and Davis[250] reported a higher response rate for a promised incentive of $2 as opposed to $1, but this difference did not reach statistical significance.

Furse and Stewart,[244] whose study focused mainly on the impact of a promised donation to charity, found a significant linear trend towards higher response rates with a larger personal incentive; this was true of all respondents combined, and on controlling for presence/absence of a charitable incentive. Examination of the RRs, however, show that the 95% CIs for the relevant individual comparisons include unity.

Fiset and colleagues[179] conducted two comparisons of the impact of $5 and $10 incentives in a survey of dentists. In the first study, only one postcard follow-up (to the entire sample) was used; in the second, two reminders, each including a duplicate questionnaire, were sent to non-respondents. In neither study was there a significant difference in response rate by incentive level, although the more intensive follow-up resulted in a significantly higher response rate. The authors also noted that a sizeable percentage of respondents (36/318; 11.3%) did not cash their cheques, more than off-setting those non-respondents who did so (22/172; 12.8%), allaying concerns about providing up-front payments while operating within budgetary constraints.

Other aspects of nature of incentive

Findings from the two studies comparing different forms of incentive provide mixed evidence for the notion that a "bird in the hand" is preferable.

Dommeyer's study[257] facilitated the comparison of three methods of providing a 25 cent enclosed incentive: coin, cheque or money order. The response rate for the coin was higher but the differences did not reach statistical significance. When all pair-wise comparisons were made, only those for the 25 cent coin versus a promised "early bird" incentive and the 25 cent coin versus a promised lottery entry reached statistical significance.

Suhre[163] found that a promised personal incentive of $10 and a promised entry in a lottery were similar in their effects on response rates.

Evidence in support of theories regarding incentives

Biner and Barton[234] sought to test whether equity theory or reactance theory came into play when an incentive was offered. In a study of factorial design they compared the effects of different values of incentive (25 cents versus $1) and of stating either that the enclosed money was intended to "obligate" the recipient to complete the questionnaire or that it was to show appreciation. The overall response rate was higher for the larger incentive, suggesting that equity theory underpinned the decision to respond.

Feedback of results

A copy of the findings from a survey may be regarded as a particular form of incentive. Dillman[1] advocated offering respondents a copy of the results, claiming that, when such an offer is made, between one-half and two-thirds of respondents take it up, and that, even in the absence of an offer, some respondents make a request for the findings. Erdos and Morgan[261] offered more qualified support for this idea, suggesting that an offer of the results increases response rates when the survey topic is of interest to the respondents. Sudman,[180] in his article on enhancing response rates to postal surveys of professionals, stated "it is always appropriate to offer to send copies of professional papers resulting from the study as they appear" and even went as far as to suggest that "it would also be a professional courtesy to indicate that computer files for the data will be available for secondary analysis"!

Identified studies

Four studies examined the impact of promised feedback (*Table 26*; see p. 173).[154,227,240,262] All four were randomised controlled trials but none was on a health-related topic. Only one showed a significant positive effect on response rates.

In a survey of students, Powers and Alderman[154] found that response rates were significantly higher when feedback was offered. There was also a significant interaction effect between the offer of feedback and the length of the questionnaire,

with feedback being of greater value when the questionnaire was longer.

Dommeyer[240] investigated the impact of offering a summary of the findings in a survey of business students. Two different questionnaires were used, a 44-item Mind Inventory Catalogue (believed to be "interesting" to the recipients) and a 55-item Tax Survey (perceived to be "uninteresting"). No significant differences were found in response rates between those who were offered a copy of the findings and those who were not made such an offer, either overall or on controlling for level of interest. However, the respondents were much more likely to request results when an explicit offer was made. Offering a copy of the results increased the costs per usable questionnaire from $0.86 to $1.02 for the "interesting" questionnaire and from $1.44 to $1.65 for the "uninteresting" questionnaire. No non-response bias with respect to a range of demographic variables was found in comparisons across all four groups (defined by offer of results and interest level) or between those who requested results and those who did not.

Subsequently, Dommeyer[262] speculated that the lack of impact of an offer of findings observed in his earlier work[240] might have been due to recipients simply skimming the covering letter and thereby missing the offer. He therefore experimented with an additional "lift letter", a second enclosure designed to grab the recipient's attention. Target respondents were randomly assigned to receive one of three letters: a standard covering letter; a covering letter offering a copy of the survey results, if requested on the return envelope or in a separate letter; and a standard covering letter plus an eye-catching lift letter offering a copy of the results (the lift letter was printed on coloured paper and stated "Read this only if you're not responding"). Response rates for usable questionnaires (at least 50% of questions completed and questionnaire returned within 3 weeks) were significantly different across the three versions. Little difference was observed between no offer of results and an offer made in the covering letter, but an offer in the lift letter produced a lower response rate. There was no difference in speed of response across the three versions, but the item non-response rate was higher when no offer of results was made. Results were most likely to be requested when the offer was made in the covering letter. Dommeyer[262] also concluded that an offer of results is likely to appeal only to those who are already interested in the topic, a group more likely to respond to a survey even in the absence of a material reward. Green and Kvidahl[227] reported no significant differences in response rates to a postal survey of schoolteachers when a summary of results was offered. In this study, personalisation of the covering letter was also manipulated. The lowest response rates were observed for non-personalised letters with an offer of results, perhaps because of a mismatch in aspects of personalisation.

Miscellaneous

Three studies investigated miscellaneous means of enhancing response rates (*Table 27*; see p. 174).[209,263,264] Two were randomised controlled trials on health-related topics.[210,264] The third was a cross-sectional study, but was included because of its novel approach.[263]

Elkind and colleagues[209] compared response rates for envelopes upon which the return address (a university) was preprinted with those on which the address was added by using a rubber stamp, for sending questionnaires to a sample of professional psychologists. No significant difference in response rates between the two types of envelope was noted.

Salvesen and Vatten,[264] prompted by an observed increase in the rate of return of questionnaires after the appearance of a newspaper article on the subject of their study, experimented with including a copy of the relevant article to those study participants (31%) who had not responded within 2 months of initial mailing. Using a Kaplan–Meier survival procedure to analyse their data, they showed that people who did not receive the article were less likely to return their questionnaire and that their response rate never caught up with those who had received a copy of the article.

Lovelock and colleagues[263] investigated the effectiveness of personal delivery of questionnaires designed for self-completion. Personal delivery and collection resulted in a response rate of 76% (but no comparison group was included). Not-at-homes accounted for 36% of all non-responses; 12.2% of all households visited remained uncontacted after two visits and had to be replaced; and apartment dwellers were over-represented in this category of non-respondents. Refusal to participate made up 39% of all non-responses; 13.3% of all households contacted (16.5% of eligible households successfully contacted) refused to participate; the incidence of refusal was higher in neighbourhoods with more residents in the 45–64 year age group and with residents in lower educational and occupational categories. Finally, 25% of all non-responses were accounted for by those who accepted a question-

naire but subsequently failed to return it (10.7% of households who agreed to participate). Lovelock and colleagues[263] concluded that personal delivery may offer some advantages but that the cost-effectiveness of this approach may be dependent on the size of the questionnaire and on the geographical spread of potential participants' addresses.

Conclusions

Caution is required in interpreting these results, in particular in comparing findings across studies, because of the heterogeneity of study populations, survey topics and factors manipulated.

Timing of survey

- Response rates do not appear to be affected by the day of posting.
- The month of posting may affect response rates, but this effect may be topic specific.

Number and relative timing of contacts

- Response rates can be increased through multiple contacts.
- Although both prenotification and follow-up contacts are effective in stimulating response rates, the latter is likely to be more powerful.

Prenotification contacts

- Prenotification is effective in increasing response rates.
- Prenotification by letter may be more effective than prenotification by telephone.
- High involvement methods of prenotification (e.g. "foot-in-door" approaches) have not been shown conclusively to improve response rates over simple prenotification; such high involvement approaches are really feasible only when telephone or personal approaches to prenotification are made.

Follow-up contacts (reminders)

- Follow-up contacts are highly effective in increasing response rates.
- There is no conclusive evidence that a "threat" of further follow-ups made in a reminder letter enhances response rates in all circumstances.
- Including a duplicate questionnaire with the first reminder does not appear to have a significant impact on response rates, but the inclusion of a replacement questionnaire with a second reminder seems to be effective.
- There is no conclusive evidence that special mailing techniques for final reminders are superior to standard mailing. Postcard reminders appear to be as effective as letters and are generally cheaper (although there may be concerns of confidentiality in health surveys).

Postage rates and types

- Findings from both primary studies and previous reviews show no consistent advantage of class of mail, or of stamped envelopes over reply-paid envelopes.

Confidentiality/anonymity

- Assurances of complete anonymity do not significantly improve response rates and may indeed have a detrimental effect.

Personalisation

- There is little conclusive evidence of the advantages of personalisation of covering letters and envelopes *per se*, but personalisation may interact with such factors as the nature of the appeal made in the covering letter and assurances of confidentiality.

Covering letters

- Traditional-style letters are more effective than novel approaches.
- There is little conclusive evidence that the characteristics of the signatory affect response rates.
- Response rates do not appear to be positively related to handwritten signatures or colour of ink.
- No one type of appeal in the covering letter offers a consistent advantage; rather, the nature of the appeal should be matched to the anticipated motivations of the recipients.

Time cues and deadlines

- A short time cue can be effective in stimulating responses.
- Specification of a deadline for responding may increase the speed of response (and thereby reduce the number of reminders needed), but it may have no effect on overall response rates.

Sponsorship

- The impact of sponsorship appears to be situation and location specific.

Saliency

- A salient (interesting, relevant and current) topic is effective in enhancing response rates.

Incentives

- Incentives are generally an effective means of increasing responses.
- Financial incentives are likely to be more effective

than non-monetary incentives of similar value.
- Enclosed incentives are more effective than promised incentives.

Feedback of results
- Offering feedback of survey results is generally not effective in stimulating response.

Miscellaneous
- Personal delivery of questionnaires for self-completion may offer some advantages but the cost-effectiveness of this approach may be situation specific.

Recommendations for practice

Relatively few studies are health related and the generalisability of findings to this field (where response rates are typically better in any case) may be limited. It must also be noted that, for many of the reviewed areas, the findings from both primary studies and from previous reviews were at best equivocal and, in some cases, contrary to expert opinion, as set out in key texts on survey design.[1,5,16,40,194] In the recommendations that follow, those derived from mixed or negative findings in previous research are highlighted.

Despite mixed findings, what is apparent is that there is no single method of enhancing response rates that is applicable in all settings. Instead, the choice of technique should be informed by consideration of the likely barriers and motivational factors for each particular survey topic and study population. The frameworks presented by Dillman[1] and by Brown and colleagues[140] form a useful basis for deliberation.

In assessing potential methods, the researcher should consider not only the likely impact on response rates but also the potential for non-response and sample composition biases, response bias and item non-response effects, as well as implications for resources of time, money, personnel and materials. The marginal benefits of intensive approaches to enhancing response rates may be outweighed by the marginal costs.

Manipulation of a single factor is unlikely to prove fruitful. Instead, researchers should consider the total "package" of: questionnaire wording; questionnaire appearance; general motivational factors (anonymity/confidentiality; personalisation; nature of appeal; other aspects of covering letter; sponsorship; saliency); mechanical and perceptual factors (timing of survey; number, timing and method of contacts; postage rates and types); and financial and other incentives.[1]

Recommendations with an evidence base from one or more high-grade primary comparative studies

General
- Consider the possibility of interactions between factors and take care to avoid apparent mismatches (e.g. a highly personalised letter combined with an assurance of total anonymity). (Recommendation based on observed interaction effects in primary studies and on expert opinion.)

Number and relative timing of contacts
- Use multiple contacts (at least one contact in addition to the initial mailing of the questionnaire). Note, however, that ethics committees may consider highly intensive contact procedures (more then three contacts) to be overly intrusive (Key L, Newcastle and North Tyneside Joint Research Ethics Committee, Newcastle upon Tyne: personal communication, 1999). (Recommendation based on evidence from primary studies, findings from previous reviews and expert opinion.)
- Consider both prenotification and follow-up contacts.
- If resources are limited, concentrate on follow-up contacts rather than prenotification. (Recommendation based on evidence from primary studies, findings from previous reviews and expert opinion.)

Prenotification
- Consider prenotification, preferably by letter, to alert target respondents to the arrival of the questionnaire. (Recommendation based on evidence from primary studies, findings from previous reviews and expert opinion.)

Follow-up (reminders)
- Use at least one reminder to non-respondents. (Recommendation based on evidence from primary studies, findings from previous reviews and expert opinion.)
- Match the appeal in the reminder letter to the perceived motivations of the study population; a consensual approach may be appropriate to some groups, while a "threat" of further follow-up may be more effective with others. (Recommendation derived from lack of consistent findings from primary studies and previous reviews; recommendation is supported by theories of respondent behaviour.)

- If initial non-response is perceived to be related to non-delivery or mislaying of the questionnaire, consider including a duplicate questionnaire with the reminder. If two or more reminders are being used, it may be appropriate to wait until the second or subsequent reminder to enclose the duplicate questionnaire. (Recommendation derived from lack of consistent findings from primary studies and previous reviews; recommendation is supported by expert opinion.)
- Choose a mode of contact for reminders that is appropriate to the survey topic and study population. Intensive techniques such as certified or recorded delivery mailing may be considered by target respondents or by ethics committees to be overly intrusive or unduly coercive (Key L, Newcastle and North Tyneside Joint Research Ethics Committee, Newcastle upon Tyne: personal communication, 1999). Although postcard reminders may be cost-effective, concerns regarding confidentiality may preclude their use in surveys on health-related topics. (Recommendation derived from lack of consistent findings from primary studies and previous reviews; recommendation is supported by expert opinion.)

Anonymity/confidentiality

- In general, total anonymity is not appropriate. Use coded (i.e. numbered and therefore identifiable) questionnaires to facilitate follow-up and record linkage. It is appropriate to be explicit in a covering letter or information sheet about how the code number will be used (i.e. to keep a check on who has responded and thereby to allow non-respondents to be followed up). (Recommendation based on evidence from primary studies and expert opinion.)

Postage rates

- For convenience, use franking rather than postage stamps for outgoing mail and use business reply envelopes for return of questionnaires. The choice between first and second class mail should involve consideration of the relative costs, the speed with which results are required, and whether it is anticipated that respondents will be aware of or influenced by the class of mail. (Recommendation derived from lack of consistent findings from primary studies and previous reviews; this recommendation is supported by the accumulated experience of the review team.)

Personalisation

- In surveys of general populations, personalisation may offer no significant advantage. However, personalisation of covering letters is likely to be appropriate if the message in the letter suggests personal knowledge of the circumstances of the recipient or uses a self-interest appeal. For example, a personalised approach may be appropriate in a survey of patients selected because they have a particular health problem. Personalisation may also be appropriate when the target respondents are in fact personally or professionally known to the sender. (Recommendation derived from lack of consistent findings from primary studies and previous reviews.)

Covering letter: style and content

- Use a traditional letter format, including headed notepaper. (Recommendation based on evidence from one primary study and on expert opinion.)
- In most circumstances, a facsimile signature is likely to be adequate. However, care should be taken to match the degree of personalisation of the signature to personalisation of the body of the letter.
- No single type of covering letter appeal is universally appropriate. Rather, the nature of the appeal made in the covering letter should be based on the perceived motivations of the study population and should be ethically sound. (Recommendation derived from lack of evidence from primary studies of any consistent advantage of one particular type of appeal; recommendation is supported by theories of respondent behaviour.)
- Consider including in the covering letter a realistic indication of the time required for completion of the questionnaire.
- Consider specifying a deadline for response, especially if a timely response is of the essence.

Sponsorship

- If ethical and practical constraints permit, choose a study sponsor appropriate to the survey topic and study population; manipulate the covering letter and return address appropriately. In surveys on health-related topics, response rates may be enhanced if the covering letter purports to come from the recipients' healthcare provider. However, consideration should be given to whether this approach may induce response bias (e.g. if patients believe their doctor is going to see their answers, they may answer differently) and to whether it is practicable (e.g. if the hospital or general practice can actually handle the dispatch and return of questionnaires). (Recommendation based on evidence from primary studies, findings from previous reviews and expert opinion.)

Saliency

- As far as possible, ensure the saliency (relevance and interest) of the survey topic to the study

population. Fortunately, surveys on health-related topics are generally perceived to be highly salient. (Recommendation based on evidence from primary studies and findings from previous reviews.)

Incentives

- If ethical and budgetary constraints allow, consider the use of enclosed financial incentives. In making the choice, the most relevant cost to consider is the projected cost per returned questionnaire: will the likely additional yield in responses outweigh the additional cost of providing the incentive? Note also that incentives are often regarded as unethical in health research and grant-awarding bodies tend to disapprove of the practice.[3] (Recommendation based on evidence from primary studies, findings from previous reviews and expert opinion.)

Recommendations derived solely from theories of respondent behaviour, previous literature reviews and/or expert opinion

Timing of survey

- If possible, avoid the month of December for conducting postal surveys. Depending on the survey topic and the study population, avoidance of the peak holiday months (July and August) may also be advisable.

Anonymity/confidentiality

- Provide appropriate assurances of confidentiality on the questionnaire itself and in the covering letter. Clarify what confidentiality means in the context of the specific survey (generally that only the research team will be able to link the numbered questionnaire to a named individual and that individual responses will not be revealed to a third party without the explicit permission of the respondent concerned).
- If totally unidentifiable questionnaires are deemed necessary, consider the use of a numbered (and therefore identifiable) postcard to be returned under separate cover. This will facilitate the use of reminders, although not record linkage.

Covering letter: style and content

- Keep the covering letter short and use language appropriate to the target recipients. If extensive or detailed information needs to be given, consider including a separate information sheet.
- Include contact details for the research organisation and ensure that all those likely to receive enquiries are adequately briefed.

Postage rates and types

- Always include a prepaid and addressed return envelope. (Recommendation based on expert opinion; no relevant studies identified.)
- Add a return address to the outside of the outgoing envelope to facilitate the return of undeliverable mail. (Recommendation based on the accumulated experience of the review team and advice from the Royal Mail; no relevant studies identified.)

Provision of feedback and results

- Providing feedback to study respondents is probably unnecessary in surveys of the general public, but it may be appropriate in surveys of health professionals. Remember to budget for the time and cost of preparing and dispatching such feedback. (Recommendation based on the experience of the review team; evidence from primary studies shows little effect, but does not relate to surveys of special populations.)

Recommendations for future research

Methods of enhancing response rates have already been extensively researched.[27] However, much of this work has been in the fields of social, educational and market research. A high priority for research, therefore, should be to examine whether techniques previously shown to enhance response rates in non-health-related surveys are also effective in stimulating responses to health surveys. Given that response rates to surveys on health-related topics are generally higher, it is possible that there may be a ceiling effect. The authors recommend that research should also focus on whether effective methods of enhancing response rates are common to health surveys of general populations, special patient or consumer groups, and health professionals.

In designing primary studies, researchers should seek to challenge expert opinion, summarised in the frameworks provided by Dillman,[1] and by Brown and colleagues,[140] and to test theories of respondent behaviour. Experimental manipulations of aspects of survey design and administration will be best carried out in a "real world" setting, "piggy-backing" the experiment on to a real survey, rather than creating an artificial situation and carrying out a survey simply for the sake of testing one or more factors hypothesised to affect response rates. In manipulating factors, care should be taken to use a realistic combination (e.g. avoid combining a high degree of personalisation with an assurance of complete anonymity). In analyses, the interaction between manipulated factors, as well as the main effects of each factor, should be examined.

Comparative studies should use multiple outcome measures; there is little point in boosting the quantity of responses (i.e. response rates) if this is at the expense of the quality of response (e.g. increased non-response bias, less complete responses, greater response bias). Moreover, more intensive approaches, although effective, may not be cost-effective; a key outcome variable should be the cost per usable questionnaire.

The priority order for further research will depend on the study population. For example, the authors believe that research into modes of contact and follow-up is particularly relevant in respect of surveys of health professionals, while studies of partial anonymity are more important in respect of patient populations, especially for surveys on sensitive topics.

Priorities for research

- Mode of contact, especially for reminders: Anecdotally, it has been suggested that telephone reminders (perhaps with an offer to complete the questionnaire as a telephone interview) may be particularly appropriate in surveys of professional groups.
- Follow-up messages: Further research is required, particularly into whether indicating either or both of (1) the importance of the individual's response or (2) response rates to date, are beneficial in stimulating response rates.
- Partial anonymity: Following Sudman's suggestion,[181] the authors recommend a comparison of identifiable (numbered) questionnaires with unidentifiable questionnaires accompanied by an identifiable postcard to be posted back separately to indicate that the questionnaire has also been returned.
- Personal delivery/collection of self-completion questionnaires: Limited evidence from one cross-sectional study[263] suggests that this may be a useful method of boosting response rates. Research into whether the potential increase in response quantity and quality outweighs the likely additional costs is recommended.
- Personalisation: Although findings from existing studies on the effects of personalisation suggest little benefit, the authors recommend testing whether a personalised letter is more effective than a form letter in situations where the target respondents are "personally" known to the researchers (as would be the case, for example, in a survey of patients by their own GP).
- Incentives: Although personal financial incentives may be regarded as unethical or inappropriate, a promised donation to charity could be more acceptable. Although evidence to date suggests that promises of untargeted charitable donations are not very effective in stimulating response in general surveys, research into whether a promised donation to a relevant charity may be effective in surveys of specific patient or consumer groups is recommended.
- Nature of appeal in covering letters: Further comparisons of "egoistic/self-interest" versus "altruistic/social utility" appeals are recommended, especially in surveys of special patient or consumer groups.
- "Foot-in-door" techniques: Evidence on the effectiveness of these techniques is mixed and further investigation of their value (given that they are resource intensive) is warranted, especially in relation to health surveys.
- Provision of information about the survey/research topic: Knowledge is required on whether providing more detailed information about the research and the means of providing that information (covering letter versus separate information sheet) has an effect on response rates. The cost per returned questionnaire would be an important outcome variable because the inclusion of extra information may have significant resource implications. No existing studies on this topic were identified.
- Provision of time cues: There is limited evidence of the effectiveness of this approach[48] and the authors therefore recommend further investigation of specification in the covering letter and/or on the questionnaire itself of the likely time required for completion.
- "Threat" of follow-up and specification of deadline for return of the questionnaire: Comparative studies are desirable on the inclusion of a statement in the covering letter accompanying the original questionnaire that indicates that reminders will be sent if the questionnaire is not returned within, for example, 2 weeks; research on specifying a deadline for response should also be carried out. Speed of response, as well as response rates, should be monitored.
- Timing of survey: In particular, efforts should be made to ascertain whether expert advice to avoid July, August and December is borne out in practice; a key outcome variable, in addition to response rates, should be speed of response.

In addition to these primary research studies, the authors also suggest that:

- In all surveys, researchers should attempt to quantify and report the extent and nature of non-response bias, and to analyse whether there

are important differences between early and late respondents.
- Reviews of the methods and results of well-designed health-related surveys (e.g.[265,266]) should be carried out as a low-cost and low-key approach to identifying good practice in the conduct of health surveys, although caution should be exercised in generalising from surveys of specific populations.
- Qualitative research, including cognitive interviewing,[134–137] should be carried out with both lay and professional groups to investigate barriers to and facilitators of participation in surveys, including motivational factors.

TABLE 15 *Timing of mailing*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Primary and secondary outcomes	Response rates	RR (95% CI)
Olivarius and Andreasen, 1995[183]	RCT	4	Health: importance of professional disciplines in treatment of disease	Physicians (Denmark)	Postal survey	***Day of mailing:*** Thursday, intended to arrive Saturday (230: 98 GPs + 132 specialists) Saturday, intended to arrive Monday (230: 102 GPs + 128 specialists)	Primary: Response rates Secondary: None	GPs: Thursday 89% Saturday 81% Specialists: Thursday 71% Saturday 76% All combined: Thursday 78% Saturday 78%	RR = 1.09 (95% CI, 0.97 to 1.23) Thurs vs. Sat (GPs) RR = 0.93 (95% CI, 0.80 to 1.08) Thurs vs. Sat (specialists) RR = 1.00 (95% CI, 0.91 to 1.10) Thurs vs. Sat (all combined)

TABLE 16 *Number and relative timing of contacts*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Primary and secondary outcomes	Response rates	RR (95% CI)
Jones and Lang, 1980[110]	RCT	4	Non-health: details of house purchase	House purchasers (USA)	Postal survey	***No. and relative timing of contacts:*** Prenotification only (975) Postnotification only (976) Pre- and postnotification (975) Prenotification was in the form of a postcard alerting the respondent to the survey, sent 4 days before the questionnaire Postnotification was in the form of postcard thanking those who had already responded and prompting non-respondents, sent 4 days after the questionnaire (Also manipulated: order of questions; covering letter (nature of appeal); sponsorship) See also *Tables 7*, 22 and *23*)	Primary: Response rates Secondary: Sample composition bias (difference in distribution of variables between entire sampling frame and respondents) Response bias (difference between reported and actual purchase prices and dates)	Prenotification only 20% Postnotification only 26% Either pre- or postnotification 23% Pre- and postnotification 31%	***No. contacts:*** RR = 1.34 (95% CI, 1.17 to 1.53) 2 vs. 1 contact RR = 1.55 (95% CI, 1.32 to 1.81) pre- and postnotification vs. prenotification only RR = 1.21 (95% CI, 1.05 to 1.39) pre- and postnotification vs. postnotification only ***Relative timing of contacts:*** RR = 1.28 (95% CI, 1.09 to 1.51) postnotification vs. prenotification

continued

TABLE 16 contd *Number and relative timing of contacts*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Primary and secondary outcomes	Response rates	RR (95% CI)
Peterson *et al.*, 1989[184]	RCT	4	Non-health: attitudes towards middle eastern countries or leisure activities	General public (USA)	Postal survey	***No. contacts:*** 1 contact (400) 2 contacts (2400) 3 contacts (4800) 4 contacts (3200) Up to 4 contacts from: prenotification; initial mailing; 1st reminder; 2nd reminder ***Relative timing of contacts:*** Prenotification only (800) Reminder(s) only (3200) (Also manipulated: prenotification; reminders (and personalisation, sponsorship, and saliency, the effects of which were not evaluated within the experimental design)) See also *Tables 17* and *18*	Primary: Response rates Secondary: Cost	1 contact 10% 2 contacts 14% 3 contacts 18% 4 contacts 22% Prenotification only 13% Reminder(s) only 16%	***No. contacts:*** RR = 1.36 (95% CI, 1.00 to 1.85) 2 vs. 1 contact RR = 1.76 (95% CI, 1.30 to 2.38) 3 vs. 1 contact RR = 2.16 (95% CI, 1.60 to 2.92) 4 vs. 1 contact RR = 1.30 (95% CI, 1.15 to 1.46) 3 vs. 2 contacts RR = 1.59 (95% CI, 1.41 to 1.79) 4 vs. 2 contacts RR = 1.23 (95% CI, 1.12 to 1.34) 4 vs. 3 contacts ***Relative timing of contacts:*** RR = 1.25 (95% CI, 1.02 to 1.53) reminder(s) only vs. prenotification only

TABLE 17 *Prenotification*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Primary and secondary outcomes	Response rates	RR (95% CI)
Kamins, 1989[187]	RCT	4	Health: attitudes towards and use of healthcare providers	Householders with telephones (USA)	Postal survey	***"Foot-in-door" methods:*** No prenotification (100) Solicitation control (151) [101] Simple foot-in-door (128) [100] Probe foot-in-door (124) [102] Labelled probe foot-in-door (119) [102] [Exclusive of initial refusals]	Primary: Response rates (initial; to follow-up; final; all calculated net of initial refusals) Secondary: None	Initial response rates: No pre-notification 31% Solicitation control 40% Simple foot 43% Probe foot 47% Labelled probe foot 59% Response to follow-up: No pre-notification 16% Solicitation control 18% Simple foot 19% Probe foot 33% Labelled probe foot 39% Final response rates: No pre-notification 41% Solicitation control 49% Simple foot 52% Probe foot 61% Labelled probe foot 72% Any pre-notification 58%	***Any prior contact vs. none:*** RR = 1.42 (95% CI, 1.11 to 1.82) any prenotification vs. none RR = 1.18 (95% CI, 0.87 to 1.61) solicitation vs. none RR = 1.27 (95% CI, 0.94 to 1.71) simple foot vs. none RR = 1.48 (95% CI, 1.12 to 1.97) probe foot vs. none RR = 1.75 (95% CI, 1.34 to 2.28) labelled probe vs. none ***Nature of prior contact:*** RR = 1.07 (95% CI, 0.81 to 1.41) simple foot vs. solicitation RR = 1.25 (95% CI, 0.97 to 1.62) probe foot vs. solicitation RR = 1.48 (95% CI, 1.17 to 1.87) labelled probe vs. solicitation RR = 1.17 (95% CI, 0.92 to 1.49) probe foot vs. simple foot RR = 1.38 (95% CI, 1.10 to 1.72) labelled probe vs. simple foot RR = 1.18 (95% CI, 0.97 to 1.44) labelled probe vs. probe foot (All results relate to final response rates)

continued

TABLE 17 contd *Prenotification*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Primary and secondary outcomes	Response rates	RR (95% CI)
Peterson *et al.*, 1989[184]	RCT	4	Non-health: attitudes towards middle eastern countries or leisure activities	General public (USA)	Postal survey	***Any prenotification vs. none:*** No prenotification (3600) Any prenotification (7200) ***Mode of prenotification:*** Postcard (3600) Letter (3600) (Also manipulated: no. contacts; reminders (and personalisation, sponsorship, and saliency, the effects of which were not evaluated within the experimental design)) See also *Tables 16* and *18*	Primary: Response rates Secondary: Cost per response	No pre-notification 15% Any pre-notification 19% Postcard 19% Letter 20%	***Any prenotification vs. none:*** RR = 1.28 (95% CI, 1.17 to 1.40) any vs. none RR = 1.25 (95% CI, 1.12 to 1.38) postcard vs. none RR = 1.31 (95% CI, 1.18 to 1.45) letter vs. none ***Mode of prenotification:*** RR = 1.05 (95% CI, 0.95 to 1.15) letter vs. postcard
Allen *et al.*, 1980[189]	RCT	4	Non-health: alienation from market place/views on global energy	Householders (Sweden)	Postal survey	***"Foot-in-door" method:*** No prior call (836) Solicitation prior call (98) Questioning prior call (98) Prior contact was by telephone	Primary: Response rates Secondary: None	No prior call: 22% Solicitation prior: 69% Questioning prior: 67% Any prior: 68%	***Any "foot-in-door" vs. no prior call:*** RR = 3.07 (95% CI, 2.62 to 3.60) any prior vs. no prior RR = 3.12 (95% CI, 2.60 to 3.74) solicitation prior vs. no prior RR = 3.03 (95% CI, 2.51 to 3.65) questioning prior vs. no prior ***Type of "foot-in-door":*** RR = 0.97 (95% CI, 0.80 to 1.17) questioning vs. solicitation prior
Hansen and Robinson, 1980[151]	RCT	4	Non-health: general public's attitudes towards most recent car purchase	Car purchasers (USA)	Postal survey	***"Foot-in-door" method:*** No prior call (200) Yes/no prior call (200) Probe prior call (200) (Also manipulated length of questionnaire) See also *Table 9*	Primary: Response rates Secondary: Speed of response Item non-response rates	No prior call 23% Yes/no prior 38% Probe prior 52% Any prior 45%	***Any "foot-in-door" vs. no prior call:*** RR = 1.91 (95% CI, 1.46 to 2.51) any prior vs. no prior RR = 1.62 (95% CI, 1.19 to 2.20) yes/no prior vs. no prior call RR = 2.23 (95% CI, 1.68 to 2.96) probe prior vs. no prior ***Type of "foot-in-door"*** RR = 1.38 (95% CI, 1.11 to 1.72) probe prior vs. yes/no prior

continued

TABLE 17 contd *Prenotification*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Primary and secondary outcomes	Response rates	RR (95% CI)
Nederhof, 1982[190]	RCT	4	Non-health: topic unspecified	General public (The Netherlands) (Sample was stratified by gender and political attitude)	Postal survey	***Mode of prenotification:*** Mail (72) Telephone (72)	Primary: Response rates Secondary: Sample composition bias	Mail 65% Telephone 68%	RR = 1.04 (95% CI, 0.83 to 1.31) telephone vs. mail
Jobber and Sanderson, 1983[164]	RCT	4	Non-health: marketing strategies	Directors of textile companies (UK)	Postal survey	***Any prenotification vs. none:*** Prior letter (400) No prior letter (400) (Also manipulated colour of paper) See also *Table 11*	Primary: Response rates Secondary: None	Initial response rates: No prior letter 41% Prior letter 39% Response to reminder: No prior letter 33% Prior letter 24% Final response rates (after 1 reminder): No prior letter 61% Prior letter 54%	Initial response rates: RR = 0.94 (95% CI, 0.79 to 1.11) prior letter vs. no prior letter Response to reminder: RR = 0.73 (95% CI, 0.55 to 0.98) prior letter vs. no prior letter Final response rates: RR = 0.88 (95% CI, 0.78 to 1.00) prior letter vs. no prior letter
Martin *et al.*, 1984,[191] 1989[192] (same survey reported in 2 articles)	RCT	4	Non-health: attitudes pertinent to planning of university services	University students (USA)	Postal survey	***Any prenotification vs. none:*** No prenotification (1000) Prenotification (1000) (Also manipulated: follow-up; postage type; personalistion) See also *Tables 18,19* and *21*	Primary: Response rates (calculated from those deliverable) Secondary: Cost per response	No prenotification 15% Prenotification 31%	RR = 2.05 (95% CI, 1.72 to 2.43) prenotification vs. no prenotification

continued

TABLE 17 contd *Prenotification*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Primary and secondary outcomes	Response rates	RR (95% CI)
Faria *et al.*, 1990[193]	RCT	4	Non-health: issues relating to home ownership	Home owners (Canada)	Postal survey	***Any prenotification vs. none:*** No prenotification (165) Any prenotification (330) ***Mode of prenotification:*** Mail – letter (165) Telephone (165)	Primary: Response rates Secondary: Speed of response Item non-response rates Cost per response	No pre-notification 34% Any pre-notification 45% Mail 48% Telephone 42%	***Any prenotification vs. none:*** RR = 1.34 (95% CI, 1.04 to 1.71) any vs. none RR = 1.42 (95% CI, 1.08 to 1.85) mail vs. none RR = 1.25 (95% CI, 0.95 to 1.66) telephone vs. none ***Mode of prenotification:*** RR = 0.88 (95% CI, 0.70 to 1.13) telephone vs. mail

TABLE 18 *Reminders*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Primary and secondary outcomes	Response rates	RR (95% CI)
Blass *et al.*, 1981[198]	RCT	4	Health: behaviour and attitudes regarding continuing education	Community psychologists (USA)	Postal survey	***Content of reminder message:*** Consensus, threat (117) Consensus, no threat (117) No consensus, threat (117) No consensus, no threat (117)	Primary: Response rates Secondary: Speed of response	Consensus, threat 37% Consensus, no threat 33% No consensus, threat 37% No consensus, no threat 28% Consensus 35% No consensus 32% Threat 37% No threat 31%	RR = 1.10 (95% CI, 0.78 to 1.56) consensus, threat vs. consensus, no threat RR = 1.30 (95% CI, 0.90 to 1.89) no consensus, threat vs. no consensus, no threat RR = 1.08 (95% CI, 0.84 to 1.39) consensus vs. no consensus RR = 1.19 (95% CI, 0.93 to 1.54) threat vs. no threat

continued

TABLE 18 contd *Reminders*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Primary and secondary outcomes	Response rates	RR (95% CI)
Roberts *et al.*, 1993[201]	RCT	4	Health: health and lifestyles	Adults on Family Health Services Authority register (UK)	Postal survey	***Mode of reminder:*** Letter (including duplicate questionnaire and return envelope) (233) Postcard (251) ***Inclusion of duplicate questionnaire:*** No duplicate questionnaire (251) Duplicate questionnaire (233) (Manipulation was in respect of the 1st reminder, sent at 3 weeks; residual non-respondents were sent a 2nd reminder in the form of a letter and duplicate questionnaire at 6 weeks)	Primary: Response rates (to 1st reminder; to 2nd reminder; to both reminders combined) Secondary: Cost per response	To 1st reminder: Letter 26% Postcard 23% To 2nd reminder: Letter 1st reminder 16% Postcard 1st reminder 28% To both reminders combined: Letter 1st reminder 37% Postcard 1st reminder 44%	To 1st reminder: RR = 1.11 (95% CI, 0.81 to 1.53) letter vs. postcard To 2nd reminder: RR = 0.57 (95% CI, 0.38 to 0.86) letter vs. postcard (at 1st reminder) To both reminders combined: RR = 0.84 (95% CI, 0.68 to 1.05) letter vs. postcard (at 1st reminder)
Nevin and Ford, 1976[195]	RCT	4	Non-health: attitudes towards halls of residence	University students (USA)	Postal survey	***Content of reminder letter:*** Casual approach (328) Veiled threat (342) (Also manipulated covering letter (specification of deadline for response)) See also *Table 22*	Primary: Response rates Secondary: Speed of response Item non-response rates Response bias	Casual approach 23% Veiled threat 38%	RR = 1.68 (95% CI, 1.32 to 2.14) veiled threat vs. casual approach

continued

TABLE 18 contd *Reminders*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Primary and secondary outcomes	Response rates	RR (95% CI)
Kahle and Sales, 1978[196]	RCT	2	Non-health: involuntary civil commitment	Psychiatrists and clinical psychologists (USA)	Postal survey	***Mode of reminder:*** Final (2nd) reminder sent by certified mail (no. not stated) Final (2nd) reminder sent by 1st class mail with air mail stickers (no. not stated) After 1st reminder, half of non-respondents were assigned to each treatment; original no. potential respondents = 880	Primary: Response rates Secondary: None	Overall final response rate 65% Original mailing replying after final reminder: Certified mailing 13% 1st class mailing 11% Undelivered letters returned after final reminder: Certified mailing 7% 1st class mailing 5%	Differences in % of original mailing replying after certified and 1st class final mailings reported as "not statistically significant" (insufficient data presented to calculate RRs and associated CIs) Reported that there were "no differences" in % of undelivered letters returned after final mailing (insufficient data presented to calculate RRs and associated CIs)
Swan *et al.*, 1980[197]	RCT	4	Non-health: educational needs	Real estate brokers (USA)	Postal survey	***Inclusion of duplicate questionnaire:*** 1st reminder: No duplicate (456) Duplicate (456) 2nd reminder: No duplicate (323) Duplicate (323) All reminders were in the form of a letter; non-respondents at each wave were randomly assigned to receive a letter only or a letter + duplicate questionnaire; thus, some will have received 2 duplicate questionnaires	Primary: Response rates Secondary: None	Response rates to 1st reminder: No duplicate 7.5% Duplicate 7.7% Response rates to 2nd reminder: No duplicate 4% Duplicate 8%	Response rates to 1st reminder: RR = 1.03 (95% CI, 0.65 to 1.62) duplicate vs. no duplicate Response rates to 2nd reminder: RR = 1.86 (95% CI, 0.99 to 3.49) duplicate vs. no duplicate

continued

TABLE 18 contd *Reminders*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Primary and secondary outcomes	Response rates	RR (95% CI)
Martin *et al.*, 1984,[191] 1989[192] (same survey reported in 2 articles)	RT	4	Non-health: attitudes pertinent to planning of university services	University students (USA)	Postal survey	***Use of a reminder:*** No reminder (1000) Reminder (1000) (Also manipulated: prenotification; postage type; personalisation) See also *Tables 17, 19* and *21*	Primary: Response rates (calculated from those deliverable) Secondary: Cost per response	No reminder 21% Reminder 25%	RR = 1.19 (95% CI, 1.02 to 1.40) reminder vs. no reminder
Dommeyer, 1987[199]	RCT	4	Non-health: attitudes towards mail surveys	Householders with telephones (USA)	Postal survey	***Nature of reminder message:*** No threat (176) Threat 1 (176) Threat 2 (176) Threat 3 (176) Threat 4 (176) 25 cents incentive (176) Threats related to speed and mode of follow-up; for the purposes of analysis the no threat and incentive groups were termed "positive" while all threat groups were termed "negative"	Primary: Response rates (75% of questions answered; denominator was no. deliverable questionnaires) Secondary: Survey withdrawals Speed of response Item non-response rates Response quality (likely validity) Group answer bias Cost per response	No threat 19% Threat 1 16% Threat 2 20% Threat 3 12% Threat 4 19% 25 cents incentive 24% "Positive" 21% "Negative" 17%	RR = 0.60 (95% CI, 0.36 to 0.99) for threat 3 vs. threat 2 RR = 2.00 (95% CI, 1.24 to 3.23) incentive vs. threat 3 95% CIs for all other comparisons include 1 RR = 1.28 (95% CI, 0.99 to 1.66) "positive" vs. "negative"

continued

TABLE 18 contd *Reminders*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Primary and secondary outcomes	Response rates	RR (95% CI)
Peterson *et al.*, 1989[184]	RCT	4	Non-health: attitudes towards middle eastern countries or leisure activities	General public (USA)	Postal survey	***Any reminders vs. none:*** No reminders (1200) 1 or 2 reminders (9600) ***No. reminders:*** None (1200) One (4800) Two (4800) ***Nature of 1st reminder:*** None (3600) Postcard (3600) Letter + questionnaire (3800) ***Nature of 2nd reminder:*** None (3600) Postcard (3600) Letter + questionnaire (3800) ***Inclusion of duplicate questionnaire with a reminder:*** With no reminders (3600) With 1 or 2 reminders (6000) (Also manipulated: prenotification; no. contacts (and personalisation, sponsorship and saliency, the effects of which were not evaluated within the experimental design)) See also *Tables 16* and *17*	Primary: Response rates Secondary: Cost per response	***Any reminders vs. none:*** No reminders 12% At least 1 reminder 19% ***No. reminders:*** 1 reminder 17% 2 reminders 20% ***Nature of 1st reminder:*** None 15% Any mode 19% Postcard 19% Letter + questionnaire 19% ***Nature of 2nd reminder:*** None 15% Any mode 19% Postcard 17% Letter + questionnaire 21% ***Inclusion of duplicate questionnaire with a reminder:*** With none 16% With 1 or 2 20%	***Any reminders vs. none:*** RR = 1.59 (95% CI, 1.35 to 1.86) ≥1 vs. none RR = 1.44 (95% CI, 1.22 to 1.70) 1 vs. none RR = 1.73 (95% CI, 1.47 to 2.04) 2 vs. none ***No. reminders:*** RR = 1.20 (95% CI, 1.11 to 1.31) 2 vs. 1 ***Nature of 1st reminder:*** RR = 1.23 (95% CI, 1.12 to 1.34) any vs. none RR = 1.24 (95% CI, 1.12 to 1.38) postcard vs. none RR = 1.21 (95% CI, 1.10 to 1.35) letter vs. none RR = 0.98 (95% CI, 0.89 to 1.08) letter vs. postcard ***Nature of 2nd reminder:*** RR = 1.30 (95% CI, 1.18 to 1.42) any vs. none RR = 1.16 (95% CI, 1.04 to 1.29) postcard vs. none RR = 1.44 (95% CI, 1.30 to 1.59) letter vs. none RR = 1.25 (95% CI, 1.13 to 1.37) letter vs. postcard ***Inclusion of duplicate questionnaire with a reminder:*** RR = 1.25 (95% CI, 1.14 to 1.37) inclusion with at least 1 reminder vs. no duplicate

continued

TABLE 18 contd *Reminders*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Primary and secondary outcomes	Response rates	RR (95% CI)
Gitelson and Drogin, 1992[200]	RCT	4	Non-health: opinions of farm show	Farm show attendees (USA)	Postal survey	***Mode of contact:*** Non-personalised letter + standard mailing (150) Personalised letter + standard mailing (150) Personalised letter + certified mailing (150)	Primary: Response rates Secondary: None	Non-personalised letter + standard mailing 13% Personalised letter + standard mailing 17% Personalised letter + certified mailing 43% Non-personalised 13% Personalised 30% Standard mailing 15% Certified mailing 43%	RR = 2.50 (95% CI, 1.69 to 3.71) personalised, certified vs. personalised, standard RR = 3.25 (95% CI, 2.08 to 5.08) personalised, certified vs. non-personalised, standard RR = 1.30 (95% CI, 0.76 to 2.22) personalised, standard vs. non-personalised, standard RR = 2.25 (95% CI, 1.44 to 3.50) personalised vs. non-personalised RR = 2.89 (95% CI, 2.09 to 4.00) certified vs. standard mailing

TABLE 19 *Postal rates and types*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Primary and secondary outcomes	Response rates	RR (95% CI)
Corcoran, 1985[208]	RCT	4	Health: clinical burn-out	Social workers (USA)	Postal survey	***Return post – stamps vs. business reply:*** Reply-permit envelopes (150) 1st class stamped reply envelopes (150)	Primary: Response rates (initial; final) Secondary: Cost of response	Initial response rates: Reply-permit 34% 1st class stamped 45% Final response rates (after 1 reminder): Reply-permit 46% 1st class stamped 50%	Initial response rates: RR = 1.33 (95% CI, 1.00 to 1.77) stamped vs. reply-permit Final response rates: RR = 1.09 (95% CI, 0.86 to 1.38) stamped vs. reply-permit

continued

TABLE 19 contd *Postal rates and types*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Primary and secondary outcomes	Response rates	RR (95% CI)
Elkind *et al.*, 1986[209]	RCT	4	Health: experience of abuse and harassment by patients	Professional psychologists (USA)	Postal survey	***Return post – stamps vs. business reply:*** Business reply return envelope (250) Stamped return envelope (250) (Also manipulated format of return address) See also *Table* 27	Primary: Response rates Secondary: None	Business reply 64.8% Stamped 64.4%	R = 0.99 (95% CI, 0.87 to 1.13) stamped vs. business reply
Cartwright and Windsor, 1989[211]	RCT	3	Health: hospital referrals and attendances	Persons on electoral register (UK)	Postal survey	***Outgoing and return post – 1st vs. 2nd class postage:*** 2nd class outgoing and reply envelope (800) 1st class outgoing and reply envelope (800) (Also manipulated covering letter (nature of appeal) See also *Table* 22	Primary: Response rates (crude prior to 1st reminder; crude prior to 2nd reminder; crude final; final usable questionnaire) Secondary: Speed of response	Crude prior to 1st reminder: 2nd class post 36% 1st class post 35% Crude prior to 2nd reminder: 2nd class post 57% 1st class post 59% Crude final: 2nd class post 70% 1st class post 72% Usable final: 2nd class post 69% 1st class post 71%	Crude prior to 1st reminder: RR = 0.97 (95% CI, 0.85 to 1.11) 1st vs. 2nd class Crude prior to 2nd reminder: RR = 1.04 (95% CI, 0.95 to 1.13) 1st vs. 2nd class Crude final: RR = 1.03 (95% CI, 0.97 to 1.10) 1st vs. 2nd class Usable final: RR = 1.03 (95% CI, 0.97 to 1.10) 1st vs. 2nd class
Harris and Guffey, 1978[204]	RCT	4	Non-health: consumer survey	Consumers (USA)	Postal survey	***Return post – stamps vs. business reply:*** Business reply return envelopes (439) Stamped return envelopes (451)	Primary: Response rates Secondary: Speed of response	Business reply 30% Stamped 36%	RR = 1.20 (95% CI, 1.00 to 1.45) stamped vs. business reply

continued

TABLE 19 contd *Postal rates and types*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Primary and secondary outcomes	Response rates	RR (95% CI)
Jones and Linda, 1978[205]	RCT	4	Non-health: organisation of conventions and meetings	Organisers of conventions and meetings (USA)	Postal survey	***Return post – stamps vs. business reply:*** Business reply return envelope (1404) Regular stamp on return envelope (1404) Commemorative stamp on return envelope (1404) (Also manipulated: cover letter appeal; sponsorship) See also *Tables 22* and *23*	Primary: Response rates Secondary: Item non-response rates Response bias	Business reply 25% Regular stamp 33% Commemorative stamp 31% Any stamp 32%	RR = 1.28 (95% CI, 1.10 to 1.82) stamped vs. reply paid RR = 1.32 (95% CI, 1.08 to 1.62) regular stamp vs. reply paid RR = 1.25 (95% CI, 1.02 to 1.54) commemorative stamp vs. reply paid RR = 0.95 (95% CI, 0.79 to 1.14) commemorative vs. regular stamp
Kahle and Sales, 1978[196]	RCT	4	Non-health: involuntary civil commitment	Psychiatrists and clinical psychologists (USA)	Postal survey	***Outgoing post – stamp vs. bulk mail:*** Bulk rate permit on outgoing envelope for reminder (200) 1st class stamp on outgoing envelope for reminder (100) (Also manipulated personalisation) See also *Table 21*	Primary: Response rates Secondary: None	Bulk mail permit 58% 1st class stamp 62%	RR = 1.07 (95% CI, 0.88 to 1.30) stamped vs. bulk rate
Labrecque, 1978[206]	RCT	4	Non-health: customer service at a marina	Customers of a marina (USA)	Postal survey	***Return post – type of stamp:*** Regular stamp on return envelope (100) Commemorative stamp on return envelope (100) (Also manipulated: personalisation; covering letter (characteristics of sender)) See also *Tables 21* and *22*	Primary: Response rates Secondary: None	Regular stamp 43% Commemorative stamp 47%	RR = 1.05 (95% CI, 0.76 to 1.44) commemorative vs. regular stamp

continued

TABLE 19 contd *Postal rates and types*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Primary and secondary outcomes	Response rates	RR (95% CI)
Hopkins and Podolak, 1983[207]	RCT	4	Non-health: views on emission control	Mechanics (USA)	Postal survey	***Outgoing post – class of mail:*** 3rd class mail (190 study 1; 170 study 2) 1st class mail (191 study 1; 169 study 2) (Also manipulated incentives) See also *Table 25*	Primary: Response rates Secondary: None	Study 1: 3rd class mail, 39% total 1st class mail, 44% total Study 2: 3rd class mail 19% 1st class mail 14%	Study 1: RR = 1.12 (95% CI, 0.88 to 1.42) 1st vs. 3rd class Study 2: RR = 0.75 (95% CI, 0.46 to 1.22) 1st vs. 3rd class
Martin *et al.*, 1984,[191] 1989[192] (same survey reported in 2 articles)	RCT	4	Non-health: attitudes pertinent to planning university services	University students (USA)	Postal survey	***Return post – stamps vs. business reply:*** Business reply envelope (1000) Stamped envelope (1000) (Also manipulated: prenotification; follow-up; personalisation) See also *Tables 17, 18* and *21*	Primary: Response rates Secondary: Cost per response	Business reply 23.5% Stamped 23.0%	RR = 0.98 (95% CI, 0.83 to 1.15) stamped vs. business reply
Harvey, 1986[210]	RCT	4	Non-health: interest and involvement in fine arts	Persons on electoral register (UK)	Postal survey	***Return post – class of mail:*** 2nd class reply envelopes (400) 1st class reply envelopes (400)	Primary: Response rates Secondary: None	2nd class reply paid 50% 1st class reply paid 48%	RR = 0.96 (95% CI, 0.83 to 1.10) 1st vs. 2nd class

TABLE 20 *Anonymity and confidentiality*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Primary and secondary outcomes	Response rates	RR (95% CI)
Campbell and Waters, 1990[215]	RCT	4	Health: knowledge of AIDS	Persons on electoral roll (UK)	Postal survey	6 separate surveys, each with: Numbered questionnaire (150) Unidentifiable questionnaire 150)	Primary: Response rates (initial; final) Secondary: None	Mean initial response rates (all 6 surveys combined): Numbered questionnaires 51% Unidentifiable questionnaires 49% Mean final response rates (after reminders; all 6 surveys combined): Numbered questionnaires 72% Unidentifiable questionnaires 49%	Mean initial response rates: RR = 0.97 (95% CI, 0.88 to 1.06) unidentifiable vs. numbered Mean final response rates: RR = 0.69 (95% CI, 0.63 to 0.74) unidentifiable vs. numbered
Jones, 1979[213]	RCT	3	Non-health: outdoor recreation behaviour	Householders with telephone and/or car (USA)	Postal survey	Anonymity not explicitly assured (11,675) Anonymity explicitly assured (11,675) (Also manipulated sponsorship) See also *Table 23*	Primary: Response rates Secondary: None	Reported that no significant main effect of anonymity; insufficient data to calculate actual response rates	Insufficient data to calculate RR
McDaniel and Rao, 1981[214]	RCT	4	Non-health: appliance warranties	Purchasers of major appliances (USA)	Postal survey	Respondent asked to sign completed questionnaire (435) Anonymity assured (435)	Primary: Response rates (based on "usable" questionnaires) Secondary: Items non-response rates No. response errors Completeness of response to open-ended question	Identifiable signed 24% Anonymous 27%	RR = 1.10 (95% CI, 0.87 to 1.40) anonymous vs. identifiable

continued

TABLE 20 contd *Anonymity and confidentiality*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Primary and secondary outcomes	Response rates	RR (95% CI)
Childers and Skinner, 1985[185]	RCT	4	Non-health: car insurance	Insurance policy holders (USA)	Postal survey	Issue of confidentiality not explicitly addressed (500) Issue of confidentiality explicitly addressed (500) (Also manipulated personalisation) See also *Table 21*	Primary: Response rates Secondary: Speed of response Completeness of response Response bias Identification of respondent	Explicit addressing of confidentiality 59% No explicit addressing of confidentiality 59%	RR = 1.00 (95% CI, 0.91 to 1.11) explicit addressing vs. no explicit addressing of confidentiality
McKee, 1992[216]	RCT	3	Non-health: issues related to programmes of a professional organisation	Members of a professional organisation (USA)	Postal survey	Numbered questionnaire, mention of follow-up (140) Un-numbered questionnaire, no mention of follow-up (140)	Primary: Response rates (initial; final) Secondary: Perceived topic involvement % Closed questions answered Completeness of responses to open-ended questions	Initial response rates: Numbered questionnaire 59% Un-numbered questionnaire 39% Final response rates: Numbered questionnaire 77% Un-numbered questionnaire 54%	Initial response rates: RR = 1.52 (95% CI, 1.18 to 1.95) numbered vs. un-numbered Final response rates: RR = 1.42 (95% CI, 1.19 to 1.70) numbered vs. un-numbered

TABLE 21 *Personalisation*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Primary and secondary outcomes	Response rates	RR (95% CI)
Roberts *et al.*, 1978[221]	RCT	4	Health: dental topics	General dental practitioners (Australia)	Postal survey	***Personalisation of letter (initial mailing):*** Form initial mailing (528) Personalised initial mailing (516) (Also manipulated covering letter (nature of appeal and specification of deadline)) See also *Table 22*	Primary: Response rates (initial; final) Secondary: None	Initial response rates: Form letter 31% Personalised letter 32% Final response rates: Form letter 69.5% Personalised letter 70.0%	Initial response rates: RR = 1.03 (95% CI, 0.95 to 1.12) personalised vs. form Final response rates: RR = 1.05 (95% CI, 0.88 to 1.26) personalised vs. form
Wunder and Wynn, 1988[226]	RCT	4	Health: satisfaction with services of health maintenance organisation	Subscribers to health maintenance organisation (USA)	Postal survey	***Addressing of envelope:*** Computer-generated label (1188) Hand-addressed (1187)	Primary: Response rates Secondary: Speed of response Item non-response rates Completeness of response to open-ended question	Computer-generated address label 31% Hand-addressed envelope 33%	RR = 1.04 (95% CI, 0.93 to 1.17) hand-addressed vs. computer-generated
Kahle and Sales, 1978[196]	RCT	4	Non-health: involuntary civil commitment	Psychiatrists and clinical psychologists (USA)	Postal survey	***Addressing of envelope:*** Address label (100) Typed address (200) (Also manipulated postal rate) See also *Table 19*	Primary: Response rates Secondary: None	Address label 52% Any typed address 63% Bulk rate and labelled 52% Stamped and typed 62% Bulk rate and typed 64%	RR = 1.21 (95% CI, 0.98 to 1.50) typed vs. labelled RR = 0.97 (95% CI, 0.78 to 1.20) stamped and typed vs. bulk rate and typed RR = 1.19 (95% CI, 0.94 to 1.52) stamped and typed vs. bulk rate and labelled

continued

TABLE 21 contd *Personalisation*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Primary and secondary outcomes	Response rates	RR (95% CI)
Labrecque, 1978[206]	RCT	4	Non-health: customer service at a marina	Customers of a marina (USA)	Postal survey	***Addressing of letter and envelope:*** Non-personalised letter and envelope (100) Personalised (hand-addressed) letter and envelope (100) (Also manipulated: postal type; covering letter (characteristics of sender)) See also *Tables 19* and *22*	Primary: Response rates Secondary: None	No personalisation 44% Personalisation 43%	RR = 0.95 (95% CI, 0.69 to 1.31) personalisation vs. no personalisation
Childers *et al.*, 1980[222]	RCT	4	Non-health: Study 1: Basic marketing text books Study 2: Specific informational publication	Study 1: Academics Study 2: Business practitioners (both USA)	Postal survey	***Personalisation of appeal:*** Study 1: Handwritten appeal (300) Typed appeal (300) Study 2: Handwritten appeal (500) Typed appeal (500) (Also manipulated covering letter (nature of appeal)) See also *Table 22*	Primary: Response rates Secondary: None	Study 1: Typed appeal 36% Handwritten appeal 33% Study 2: Typed appeal 31% Handwritten appeal 34%	Study 1 (academic sample): RR = 0.92 (95% CI, 0.73 to 1.16) handwritten vs. typed Study 2 (business sample): RR = 1.10 (95% CI, 0.91 to 1.33) handwritten vs. typed
Martin *et al.*, 1984,[191] 1989[192] (same survey reported in 2 articles)	RCT	4	Non-health: attitudes pertinent to planning university services	University students (USA)	Postal survey	***Personalisation of letter:*** Non-personalised letter (1000) Personalised letter (1000) (Also manipulated: prenotification; follow-up; postage type) See also *Tables 17–19*	Primary: Response rates Secondary: Cost per response	Non-personalised letter 15% Personalised letter 24%	RR = 1.10 (95% CI, 0.93 to 1.28) personalisation vs. no personalisation

continued

TABLE 21 contd *Personalisation*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Primary and secondary outcomes	Response rates	RR (95% CI)
Childers and Skinner, 1985[185]	RCT	4	Non-health: car insurance	Insurance policy holders (USA)	Postal survey	***Personalisation of envelopes (outgoing and return):*** Computer-printed address on outgoing envelope (500) Labelled address on outgoing envelope (500) Computer-printed address on return envelope (500) Labelled address on return envelope (500) This study used a factorial design (Also manipulated anonymity/confidentiality) See also *Table 20*	Primary: Response rates Secondary: Speed of response Completeness of response Response bias Identification of respondent	Labelled address on outgoing envelope 58% Computer-printed address on outgoing envelope 62% Labelled address on return envelope 58% Computer-printed address on return envelope 62%	RR =1.07 (95% CI, 0.96 to 1.18) computer-printed vs. labelled (outgoing) RR = 1.07 (95% CI, 0.96 to 1.18) computer-printed vs. labelled (return)
Woodward and McKelvie, 1985[223]	RCT	4	Non-health: attitudes towards trades unions + attitudes towards punishment of criminals	University students of business and social science (USA)	Postal survey	***Degree of personalisation:*** Addressed to student box no. alone (100) Addressed to box no. + "Dear Mr/Ms Surname" (100) Addressed to box no. + "Dear Forename Surname" (100) Addressed to box no. + "Dear Nickname Surname" (100) Nickname was a shortened version of the forename, e.g. Joe for Joseph) (Also manipulated saliency) See also *Table 24*	Primary: Response rates Secondary: None	Box no. alone 26% Surname 31% Forename/ surname 29% Nickname/ surname 41%	RR = 1.19 (95% CI, 0.77 to 1.85) surname vs. box no. RR = 1.12 (95% CI, 0.71 to 1.75) forename/surname vs. box no. RR = 1.58 (95% CI, 1.05 to 2.37) nickname/surname vs. box no. RR = 0.94 (95% CI, 0.61 to 1.43) forename/surname vs. surname RR = 1.32 (95% CI, 0.91 to 1.92) nickname/surname vs. surname RR = 1.41 (95% CI, 0.96 to 2.08) nickname/surname vs. forename/surname

continued

TABLE 21 contd *Personalisation*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Primary and secondary outcomes	Response rates	RR (95% CI)
Worthen and Valcarce, 1985[224]	RCT	4	Non-health: opinions on college courses	School teachers (USA)	Postal survey	***Personalisation of letter (initial mailing and follow-up):*** Initial mailing: Form (500) [487] Personalised (500) [489] Follow-up: Form (365) [357] Personalised (365) [365] [Nos deliverable]	Primary: Response rates (initial; to follow-up; final; in both cases calculated on the basis of those deliverable) Secondary: None	Initial mailing: Form letter 23% Personalised letter 28% Overall response rates to follow-up mailing: Form letter 29% Personalised letter 33% Response rates to follow-up for those with personalised initial letter: Form letter 27% Personalised letter 36% Response rates to follow-up for those with form initial letter: Form letter 30.4% Personalised letter 29.8%	Initial mailing: RR = 1.21 (95% CI, 0.97 to 1.51) personalised vs. form Follow-up mailing: RR = 1.14 (95% CI, 0.92 to 1.42) personalised vs. form Response rates to follow-up for those with personalised initial letter: RR = 1.34 (95% CI, 0.97 to 1.84) personalised vs. form Response rates to follow-up for those with form initial letter: RR = 0.98 (95% CI, 0.72 to 1.33) personalised vs. form
Green and Stager, 1986[225]	RCT	3	Non-health: experience of and attitudes towards classroom testing	School teachers (USA)	Postal survey	***Personalisation of salutation and signature:*** Personalised salutation (375) Duplicated salutation (375) Hand signature (383) Facsimile signature (367)	Primary: Response rates Secondary: None	Insufficient data presented to report response rates by level of personalisation; reported that "response rates ... were essentially equivalent for personalisation and no personalisation"	Insufficient data to calculate RRs

continued

TABLE 21 contd *Personalisation*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Primary and secondary outcomes	Response rates	RR (95% CI)
Green and Kvidahl, 1989[227]	RCT	4	Non-health: opinions on applying research findings to teaching	School teachers (USA)	Postal survey	***Personalisation of letter:*** Form letter (300) Personalised letter (300) (Also manipulated feedback of results) See also *Table 26*	Primary: Response rates Secondary: Speed of response Cost per response	Form letter 66% Personalised letter 75%	RR = 1.17 (95% CI, 1.05 to 1.29) for personalised vs. form

TABLE 22 *Covering letter*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Primary and secondary outcomes	Response rates	RR (95% CI)
Roberts *et al.*, 1978[221]	RCT	4	Health: dental topics	General dental practitioners (Australia)	Postal survey	***Nature of appeal:*** No social appeal (525) Social appeal (519) ***Specification of return deadline:*** No deadline specified (517) Deadline specified (500) This study used a factorial design (Also manipulated personalisation (of covering letter)) See also *Table 21*	Primary: Response rates (initial; final) Secondary: Speed of response Cost of follow-up	***Nature of appeal:*** Initial response rates: No social appeal 33% Social appeal 30% Final response rates: No social appeal 70% Social appeal 68% ***Specification of deadline:*** Initial response rates: No deadline 28% Deadline 35% Final response rates: No deadline 68% Deadline 70%	***Nature of appeal:*** Initial response rates: RR = 0.91 (95% CI, 0.76 to 1.08) social vs. no social Final response rates: RR = 0.98 (95% CI, 0.90 to 1.06) social vs. no social ***Specification of deadline:*** Initial response rates: RR = 1.24 (95% CI, 1.03 to 1.49) deadline vs. no deadline Final response rates: RR = 1.04 (95% CI, 0.96 to 1.13) deadline vs. no deadline

continued

TABLE 22 contd *Covering letter*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Primary and secondary outcomes	Response rates	RR (95% CI)
Salomone and Miller, 1978[228]	RCT	4	Health: vocational development and job satisfaction	Rehabilitation counsellors (USA)	Postal survey	***Nature of appeal:*** Professionalism (330) Humour (330) Importance of respondent (330) Token compensation (incentive) (330) (Also manipulated incentive) See also *Table 25*	Primary: Response rates (initial; final) Secondary: None	Initial response rates: Professionalism 48% Humour 52% Importance of respondent 53% Token compensation 65% Final response rates: Professionalism 80% Humour 80% Importance of respondent 76% Token compensation 85%	Initial response rates: RR = 1.06 (95% CI, 0.91 to 1.24) humour vs. professionalism RR = 1.08 (95% CI, 0.93 to 1.26) importance of respondent vs. professionalism RR = 1.35 (95% CI, 1.18 to 1.55) token compensation vs. professionalism RR = 1.02 (95% CI, 0.88 to 1.18) importance of respondent vs. humour RR = 1.27 (95% CI, 1.11 to 1.45) token compensation vs. humour RR = 1.25 (95% CI, 1.10 to 1.42) token compensation vs. importance of respondent Final response rates: RR = 1.00 (95% CI, 0.93 to 1.08) humour vs. professionalism RR = 0.95 (95% CI, 0.88 to 1.04) importance of respondent vs. professionalism RR = 1.06 (95% CI, 0.99 to 1.14) token compensation vs. professionalism RR = 0.95 (95% CI, 0.88 to 1.04) importance of respondent vs. humour RR = 1.06 (95% CI, 0.99 to 1.14) token compensation vs. humour RR = 1.12 (95% CI, 1.03 to 1.20) token compensation vs. importance of respondent
McKillip and Lockhart, 1984[229]	RCT	4	Health: evaluation of alcohol education programme	University students (USA)	Postal survey	***Nature of appeal:*** Study 1: Value-expression (100: 100 UG) Knowledge (200: 100 UG + 100 PG) Utility (500: 400 UG + 100 PG) Study 2: Utility: (475: 400 UG + 75 PG) Combined: (200: 100 UG + 100 PG)	Primary: Response rates Secondary: None	Study 1 – UG: Value-expression 39% Knowledge 44% Utility 48% Study 1 – PG: Knowledge 57% Utility 54% Study 2 – UG: Utility 54% Combined 43% Sudy 2 – PG: Utility 55% Combined 57%	Study 1: RR = 1.24 (95% CI, 0.95 to 1.61) utility vs. value-expression RR = 1.13 (95% CI, 0.81 to 1.57) knowledge vs. value-expression RR = 1.10 (95% CI, 0.86 to 1.40) utility vs. knowledge (UG) RR = 0.95 (95% CI, 0.74 to 1.21) utility vs. knowledge (PG) Study 2: RR = 0.80 (95% CI, 0.63 to 1.03) combined vs. utility (UG) RR = 1.13 (95% CI, 0.88 to 1.47) combined vs. utility (PG)

continued

TABLE 22 contd *Covering letter*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Primary and secondary outcomes	Response rates	RR (95% CI)
Dodd *et al.*, 1988[233]	RCT	3	Health: desired and actual hospital privileges	Professional psychologists (USA)	Postal survey	***Use of postscript appeal:*** No postscript appeal (531) Postscript appeal (532) ***Colour of ink for signature:*** Blue ink (531) Bright green ink (532) ***No. signatories:*** 1 sender (531) 3 senders (532) This study used a factorial design; total sample size = 1063	Primary: Response rates Secondary: None	Overall response rate 61% No significant effect of postscript appeal (data not reported) No significant effect of colour of ink (data not reported) No significant effect of no. signatories (data not reported)	Insufficient data to calculate RR
Cartwright and Windsor, 1989[211]	RCT	3	Health: hospital referrals and attendances	Persons on electoral register (UK)	Postal survey	***Nature of appeal:*** Not asked to help in further research or to provide telephone no. (800) Asked to help in further research and to provide telephone no. (800) (Also manipulated postage rates) See also *Table 19*	Primary: Response rates Secondary: Speed of response	Not asked for help 74% Asked for help 65%	RR = 0.86 (95% CI, 0.80 to 0.91) for asked vs. not asked
Nevin and Ford, 1976[195]	RCT	4	Non-health: attitudes towards halls of residence	University students (USA)	Postal survey	***Specification of deadline:*** No deadline (255) 5-day deadline (279) 7-day deadline (262) 9-day deadline (279) (Also manipulated follow-up (content of reminder message)) See also *Table 18*	Primary: Response rates Secondary: Speed of response Item non-response rates Response bias	No deadline 50% 5-day deadline 43% 7-day deadline 49% 9-day deadline 53% Any deadline 48%	RR = 0.86 (95% CI, 0.72 to 1.04) 5-day vs. none RR = 0.97 (95% CI, 0.82 to 1.16) 7-day vs. none RR = 1.07 (95% CI, 0.91 to 1.26) 9-day vs. none RR = 1.13 (95% CI, 0.94 to 1.35) 7-day vs. 5-day RR = 1.24 (95% CI, 1.04 to 1.48) 9-day vs. 5-day RR = 1.10 (95% CI, 0.93 to 1.30) 9-day vs. 7-day RR = 0.97 (95% CI, 0.84 to 1.12) any vs. none

continued

TABLE 22 contd *Covering letter*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Primary and secondary outcomes	Response rates	RR (95% CI)
Jones and Linda, 1978[205]	RCT	4	Non-health: organisation of conventions and meetings	Organisers of conventions and meetings (USA)	Postal survey	***Nature of appeal:*** Value to sponsor (1404) Self-interest (1404) Social utility (1404) (Also manipulated: postage type; sponsorship) See also *Tables 19* and *23*	Primary: Response rates Secondary: Item non-response rate Response bias	Value to sponsor 26% Self-interest 31% Social utility 31%	RR = 1.18 (95% CI, 0.96 to 1.44) social utility vs. sponsor RR = 1.18 (95% CI, 0.96 to 1.44) self-interest vs. sponsor RR = 1.00 (95% CI, 0.83 to 1.21) social utility vs. sponsor
Labrecque, 1978[206]	RCT	4	Non-health: customer service at a marina	Customers of a marina (USA)	Postal survey	***Status of sender:*** Marina service manager (100) Marina owner (100) (Also manipulated: postage type; personalisation) See also *Tables 19* and *21*	Primary: Response rates Secondary: None	Service manager 37% Owner 51%	RR = 1.39 (95% CI, 1.00 to 1.93) owner vs. service manager
Childers *et al.*, 1980[222]	RCT	4	Non-health: Study 1: Use of and attitudes to basic marketing of text books Study 2: Attitudes to specific informational publication	Study 1: Academics Study 2: Business practitioners (USA)	Postal survey	***Nature of appeal:*** Study 1: No appeal control (100) [94] Egoistic (200) [186] Help the sponsor (200) [172] Social utility (200) [187] Study 2: No appeal (143) Egoistic (286) Help the sponsor (286) Social utility (286) [No. deliverable and eligible] (Also manipulated personalisation) See also *Table 21*	Primary: Response rates (omitting those undeliverable or ineligible) Secondary: Response completeness Response bias	Study 1: No appeal control: 44% Egoistic: 39% Help the sponsor: 38% Social utility: 28% Any appeal: 35% Study 2: No appeal control: 31% Egoistic: 31% Help the sponsor: 34% Social utility: 33% Any appeal: 33%	Study 1 (academic sample): RR = 0.89 (95% CI, 0.66 to 1.19) egoistic vs. no appeal RR = 0.87 (95% CI, 0.64 to 1.17) help sponsor vs. no appeal RR = 0.64 (95% CI, 0.46 to 0.88) social utility vs. no appeal RR = 0.98 (95% CI, 0.75 to 1.27) help sponsor vs. egoistic RR = 0.72 (95% CI, 0.54 to 0.96) social utility vs. egoistic RR = 0.74 (95% CI, 0.55 to 0.99) social utility vs. help sponsor RR = 0.80 (95% CI, 0.61 to 1.03) any appeal vs. no appeal Study 2 (business sample): RR = 1.00 (95% CI, 0.74 to 1.35) egoistic vs. no appeal RR = 1.11 (95% CI, 0.83 to 1.49) help sponsor vs. no appeal RR = 1.06 (95% CI, 0.79 to 1.42) social utility vs. no appeal RR = 1.11 (95% CI, 0.88 to 1.41) help sponsor vs. egoistic RR = 1.06 (95% CI, 0.83 to 1.34) social utility vs. egoistic RR = 0.95 (95% CI, 0.75 to 1.20) social utility vs. help sponsor RR = 1.06 (95% CI, 0.81 to 1.38) any appeal vs. no appeal

continued

TABLE 22 contd *Covering letter*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Primary and secondary outcomes	Response rates	RR (95% CI)
Jones and Lang, 1980[110]	RCT	4	Non-health: details of house purchase	House purchasers (USA)	Postal survey	***Nature of appeal:*** Egoistic (1463) Social utility (1463) (Also manipulated: order of questions; no. and timing of contacts; sponsorship) See also Tables *7*, *16* and *23*	Primary: Response rates Secondary: Sample composition bias (difference in distribution of variables between entire sampling frame and respondents) Response bias (difference between reported and actual purchase prices and dates)	Egoistic 25% Social utility 26%	RR = 1.08 (95% CI, 0.95 to 1.22) social utility vs. egoistic
Hornik, 1981[48]	RCT	4	Non-health: attitudes to TV advertising	TV viewers (USA)	Postal survey	***Provision of time cue:*** No time cue (200) Short time cue (200) Long time cue (200)	Primary: Response rates Secondary: Perceived time to complete Speed of response Item non-response rates Response bias	No time cue 32% Short time cue 42% Long time cue 26%	RR = 1.32 (95% CI, 1.01 to 1.71) short vs. none RR = 0.81 (95% CI, 0.59 to 1.11) long vs. none RR = 0.61 (95% CI, 0.46 to 0.82) long vs. short
Wagner and O'Toole, 1985[230]	RCT	4	Non-health: willingness to administer survey to students	Academic psychologists (USA)	Postal survey	***Style of letter:*** Traditional letter (53) Humorous letter (53)	Primary: Return of form regarding participation in study Secondary: None	Traditional 83% Humorous 13%	RR = 0.16 (95% CI, 0.08 to 0.32) humorous vs. traditional

continued

TABLE 22 contd *Covering letter*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Primary and secondary outcomes	Response rates	RR (95% CI)
Dodd and Markwiese, 1987[231]	RCT	4	Non-health: topic not specified	University employees (USA)	Postal survey	***Status of sender:*** Sent by student (100) Sent by professor (100) ***Gender of sender:*** Male sender (100) Female sender (100) ***Style of signature:*** Facsimile (100) Handwritten (100) This study used a factorial design; total sample size = 200	Primary: Response rates Secondary: Questionnaire completion rates	***Status of sender:*** No significant effect (actual response rates not stated) ***Gender of sender:*** No significant effect (actual response rates not stated) ***Style of signature:*** Questionnaire return rates: Facsimile 36% Handwritten 44%	RR = 1.22 (95% CI, 0.87 to 1.72) handwritten vs. facsimile
Biner, 1988[232]	RCT	4	Non-health: assessment of community needs	Householders with telephone (USA)	Postal survey	***Nature of appeal:*** Stressing personal freedom of choice (100) Likely to induce reactance (100) (Also manipulated incentives (inclusion of incentive)) See also *Table 25*	Primary: Response rates Secondary: None	Stressing personal freedom of choice 59% Reactance-inducing 44%	RR = 0.75 (95% CI, 0.56 to 1.00) reactance-inducing vs. stressing personal freedom
Biner and Barton, 1990[234]	RCT	4	Non-health: assessment of community needs	Householders with telephone (USA)	Postal survey	***Portrayal of incentive:*** As token of appreciation (100) As inducement of obligation (100) (Also manipulated incentives (size of incentive)) See also *Table 25*	Primary: Response rates Secondary: Pattern of responses	Token of appreciation 44% Inducement of obligation 66%	RR = 1.25 (95% CI, 0.99 to 1.59) inducement of obligation vs. token of appreciation

UG, undergraduate; PG, postgraduate

TABLE 23 *Sponsorship*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Primary and secondary outcomes	Response rates	RR (95% CI)
Jacoby, 1990[150]	RCT	4	Health: users' views of GP services	Persons on electoral roll (UK)	Postal survey	Research unit sponsorship (1563) FPC sponsorship (383) (Also manipulated length of questionnaire) See also *Table 9*	Primary: Response rates Secondary: Speed of response (% responding within 3 weeks) Response bias	Research unit 66% FPC 84%	RR = 1.27 (95% CI, 1.20 to 1.35) FPC vs. research unit
Smith *et al.*, 1985[239]	Non-random concurrent controlled study	3	Health: coronary heart disease	Primary care patients (UK)	Postal survey	Covering letter from research unit doctor (203) Covering letter from patient's GP (206)	Primary: Response rates (initial; final; crude and adjusted by omitting those returned as undeliverable) Secondary: None	Initial crude response rate: Research unit doctor 58% GP 67% Initial adjusted response rate: Research unit doctor 66% GP 70% Final crude response rate: Research unit doctor 65% GP 81% Final adjusted response rate: Research unit doctor 75% GP 85%	Initial crude response rate: RR = 1.16 (95% CI, 1.00 to 1.35) GP vs. research unit Initial adjusted response rate: RR = 1.07 (95% CI, 0.93 to 1.23) GP vs. research unit Final crude response rate: RR = 1.24 (95% CI, 1.10 to 1.40) GP vs. research unit Final adjusted response rate: RR = 1.13 (95% CI, 1.02 to 1.25) GP vs. research unit
Jones and Linda, 1978[206]	RCT	4	Non-health: organisation of conventions and meetings	Organisers of conventions and meetings (USA)	Postal survey	Private business (1404) Government agency (1404) University (1404) Private business was a market research organisation (Also manipulated: postage type; covering letter (nature of appeal)) See also *Tables 19* and *22*	Primary: Response rates Secondary: Item non-response rates Response bias	Private business 25% Government agency 29% University 35%	RR = 1.17 (95% CI, 0.95 to 1.45) government agency vs. private business RR = 1.40 (95% CI, 1.14 to 1.71) university vs. private business RR = 1.19 (95% CI, 0.99 to 1.44) university vs. government agency

continued

TABLE 23 contd *Sponsorship*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Primary and secondary outcomes	Response rates	RR (95% CI)
Jones, 1979[214]	RCT	3	Non-health: outdoor recreation behaviour	Householders with telephone and/or car (USA)	Postal survey	Government (11,675) University (11,675) (Also manipulated anonymity/confidentiality) See also *Table 20*	Primary: Response rates Secondary: None	Significant main effect of sponsorship ($p < 0.01$); actual values not stated and insufficient data to calculate response rates	Insufficient data to calculate RR
Jones and Lang, 1980[110]	RCT	4	Non-health: details of house purchase	House purchasers (USA)	Postal survey	Private agency (1463) University (1463) "Private agency" in this case was the local board of realtors' (estate agents) trade association (Also manipulated: order of questions; no. and timing of contacts; covering letter (nature of appeal)) See also *Tables 7, 16* and *22*	Primary: Response rates Secondary: Sample composition bias (difference in distribution of variables between entire sampling frame and respondents) Response bias (difference between reported and actual purchase prices and dates)	Private agency 22% University 29%	RR = 1.30 (95% CI, 1.15 to 1.47) for university vs. private agency

TABLE 24 *Saliency/subject matter*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Primary and secondary outcomes	Response rates	RR (95% CI)
Hovland *et al.*, 1980[179]	RCT	3	Health: dentists' attitudes and knowledge	Dentists (USA)	Postal survey	Knowledge questionnaire (200) Attitude questionnaire (200)	Primary: Response rates (initial; final) Secondary: Non-response bias	Initial response rates: Knowledge 43% Attitude 54% Final response rates: Knowledge 98% Attitude 95%	Initial response rates: RR = 1.24 (95% CI, 1.01 to 1.53) attitude vs. knowledge Final response rates: RR = 0.97 (95% CI, 0.94 to 1.01) attitude vs. knowledge
Dommeyer, 1985[240]	RCT	4	Non-health: Mind Inventory Catalogue and Tax Survey (one or other of these scales)	Business students (USA)	Postal survey	Uninteresting topic – Tax Survey (210) Interesting topic – Mind Inventory Catalogue (210) (Also manipulated feedback of results) See also *Table 26*	Primary: Response rates Secondary: Item non-response rates Speed of response Request for feedback	Uninteresting topic 31% Interesting topic 54%	RR = 1.74 (95% CI, 1.37 to 2.20) interesting vs. uninteresting
Woodward and McKelvie, 1985[223]	RCT	4	Non-health: attitudes towards trades unions + attitudes towards punishment of criminals	University students of business and social science (USA)	Postal survey	Low-interest topic (200) High-interest topic (200) 2 questionnaires, 1 on trade unions and 1 on punishment of criminals; the former was deemed to be "more interesting" to business students and "less interesting" to social science students; the latter was judged to be "more interesting" to social science students and "less interesting" to business students (Also manipulated personalisation) See also *Table 21*	Primary: Response rates Secondary: None	Low-interest topic 34% High-interest topic 30% (Interaction of division (business or social science) with interest was stated to be ns)	RR = 0.90 (95% CI, 0.77 to 1.04) high vs. low interest

TABLE 25 *Incentives*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Primary and secondary outcomes	Response rates	RR (95% CI)
Salomone and Miller, 1978[228]	RCT	4	Health: vocational development and job satisfaction	Rehabilitation counsellors (USA)	Postal survey	***Incentive (enclosed) vs. no incentive:*** No incentive (990) 25 cents enclosed (330) (Also manipulated covering letter (style)) See also *Table* 22	Primary: Response rates (initial; final) Secondary: None	Initial response rates: No incentive 51% 25 cents enclosed 65% Final response rates: No incentive 78% 25 cents enclosed 85%	Initial response rates: RR = 1.29 (95% CI, 1.17 to 1.42) 25 cents enclosed vs. no incentive Final response rates: RR = 1.08 (95% CI, 1.02 to 1.14) 25 cents enclosed vs. no incentive
Cook *et al.*, 1985[252]	RCT	4	Health: nature of drug education programme	Administrators of drug education programmes (USA)	Postal survey	***Incentive (promised) vs. no incentive:*** No incentive (125) $100 promised (125)	Primary: Response rates Secondary: Willingness to participate in further study	No incentive 27% $100 promised 22%	RR = 0.82 (95% CI, 0.53 to 1.27) $100 promised vs. none
Mortagy *et al.*, 1985[253]	RCT	4	Health: respiratory symptoms in adults	Persons on electoral register (UK)	Postal survey	***Incentive (promised) vs. no incentive:*** No incentive (950) [900] Promised entry in lottery for prizes of £50, £30 and £20 (1762) [1642] [No. deliverable]	Primary: Response rates (initial; final; in both cases omitting those undeliverable) Secondary: None	Initial response rates: No incentive 54% Promised lottery entry 57% Final response rates: No incentive 72% Promised lottery entry 73%	Initial response rates: RR = 1.04 (95% CI, 0.97 to 1.12) promised lottery vs. no incentive Final response rates: RR = 1.02 (95% CI, 0.97 to 1.07) promised lottery vs. no incentive
Woodward *et al.*, 1985[254]	RCT	4	Health: respiratory illness in childhood	Householders with children (Australia)	Postal survey	***Incentive (promised) vs. no incentive:*** No incentive (100) Promised entry in lottery for restaurant meal to value of A$100 (100)	Primary: Response rates Secondary: None	No incentive 60% Promised lottery entry 73%	RR = 1.22 (95% CI, 1.00 to 1.49) promised lottery vs. no incentive

continued

TABLE 25 contd *Incentives*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Primary and secondary outcomes	Response rates	RR (95% CI)
Weltzien *et al.*, 1986[256]	RCT	4	Health: satisfaction with mental health care services	Ex-clients of mental health centre (USA)	Postal survey	***Incentive (enclosed) vs. no incentive:*** No incentive (471) 2 cents enclosed (471)	Primary: Response rates Secondary: None	No incentive 18% 2 cents enclosed 24%	RR = 1.34 (95% CI, 1.04 to 1.71) 2 cents enclosed vs. no incentive
Berry and Kanouse, 1987[246]	RCT	4	Health: familiarity with National Institutes of Health consensus programme	Community physicians (USA)	Postal survey	***Promised vs. enclosed incentive:*** Promised $20 (1073) Enclosed $20 (1074)	Primary: Response rates Secondary: Refusal rates Non-response bias Cost per response	Promised 66% Enclosed 78%	RR = 1.18 (95% CI, 1.12 to 1.25) enclosed vs. promised
Fiset *et al.*, 1994[179]	RCT	4	Health: experience of dental malpractice liability	Dentists (USA)	Postal survey	***Size of incentive:*** $5 enclosed (100; study 1) and (159; study 2) $10 enclosed (100; study 1) and (158; study 2)	Primary: Response rates (final, omitting those who could not be located) Secondary: None	Study 1: $5 enclosed 55% $10 enclosed 61% Study 2: $5 enclosed 69.8% $10 enclosed 69.3%	Study 1: RR = 1.09 (95% CI, 0.86 to 1.39) $10 vs. $5 Study 2: RR = 0.99 (95% CI, 0.85 to 1.15) $10 vs. $5
Little and Davis, 1984[250]	Non-random concurrent controlled study	4	Health: drinking, smoking and dietary habits of women postpartum	Postpartum women (USA)	Postal survey	***Incentive vs. no incentive:*** No incentive (343) Incentive (887) ***Promised vs. enclosed incentive:*** Promised (679) Enclosed (208) ***Size of promised incentive:*** $1 promised (344) $2 promised (335) $1 enclosed (208)	Primary: Response rates Secondary: Cost per response	No incentive 59% Incentive 69% Promised 66% Enclosed 79% $1 promised 63% $2 promised 70% $1 enclosed 79%	***Incentive vs. no incentive:*** RR = 1.17 (95% CI, 1.06 to 1.29) any incentive vs. none RR = 1.07 (95% CI, 0.95 to 1.20) $1 promised vs. none RR = 1.17 (95% CI, 1.05 to 1.31) $2 promised vs. none RR = 1.33 (95% CI, 1.19 to 1.49) $1 enclosed vs. none ***Promised vs. enclosed incentive:*** RR = 1.19 (95% CI, 1.09 to 1.30) any enclosed vs. promised RR = 1.25 (95% CI, 1.13 to 1.39) $1 enclosed vs. $1 promised RR = 1.14 (95% CI, 1.03 to 1.26) $1 enclosed vs. $2 promised ***Size of promised incentive:*** RR = 1.10 (95% CI, 0.99 to 1.22) $2 promised vs. $1 promised

continued

TABLE 25 contd *Incentives*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Primary and secondary outcomes	Response rates	RR (95% CI)
Hansen, 1980[241]	RCT	4	Non-health: evaluation of hard hats	Industrial safety engineers (USA)	Postal survey	***Incentive (enclosed) vs. no incentive:*** No incentive (811) Incentive enclosed (1614) ***Financial vs. non-monetary:*** Ball-point pen enclosed (804) 25 cents enclosed (810)	Primary: Response rates Secondary: Speed of response Item non-response rates (open-ended questions) % Suggestions made in responses to open-ended questions judged to be of high quality	No incentive 14% Incentive enclosed 30% Pen enclosed 22% 25 cents enclosed 38%	***Incentive enclosed vs. none:*** RR = 2.14 (95% CI, 1.78 to 2.57) incentive enclosed vs. none RR = 1.57 (95% CI, 1.26 to 1.94) pen enclosed vs. none RR = 2.71 (95% CI, 2.23 to 3.28) 25 cents enclosed vs. none ***Nature of incentive:*** RR = 1.73 (95% CI, 1.48 to 2.02) 25 cents vs. pen
Whitmore, 1976[249]	RCT	4	Non-health: car purchasing behaviour	Car purchasers (USA)	Postal survey	***Incentive (enclosed) vs. no incentive:*** No incentive (500) Key ring enclosed (500)	Primary: Response rates Secondary: Response bias (association between response patterns for individual items and provision of incentive)	No incentive 52% Key ring enclosed 57%	RR = 1.10 (95% CI, 0.98 to 1.23) key ring enclosed vs. none

continued

TABLE 25 contd *Incentives*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Primary and secondary outcomes	Response rates	RR (95% CI)
Furse and Stewart, 1982[244]	RCT	4	Non-health: use of microwave ovens	Microwave oven users (USA)	Postal survey	***Incentive (promised and/or enclosed) vs. no incentive:*** No incentive (100) Incentive (500) ***Nature of incentive:*** 50 cents enclosed (100) $1 enclosed (100) $1 donation promised (100) 50 cents enclosed and $1 donation promised (100) $1 enclosed and $1 donation promised (100)	Primary: Response rates Secondary: Non-response bias Item non-response rates Speed of response Cost per response	No incentive 54% Incentive 70% 50 cents enclosed 68% $1 enclosed 76% $1 donation promised 56% 50 cents enclosed and $1 donation promised 71% $1 enclosed and $1 donation promised 78%	***Incentive vs. no incentive:*** RR = 1.29 (95% CI, 1.07 to 1.56) any incentive vs. none RR = 1.26 (95% CI, 1.01 to 1.58) 50 cents enclosed vs. none RR = 1.41 (95% CI, 1.14 to 1.74) $1 enclosed vs. none RR = 1.04 (95% CI, 0.81 to 1.33) $1 donation promised vs. none RR = 1.31 (95% CI, 1.06 to 1.64) 50 cents enclosed and $1 donation promised vs. none RR = 1.44 (95% CI, 1.17 to 1.78) $1 enclosed and $1 donation promised vs. none ***Nature of incentive:*** RR = 1.12 (95% CI, 0.94 to 1.33) $1 enclosed vs. 50 cents enclosed (no promised donation) RR = 0.82 (95% CI, 0.66 to 1.03) $1 promised donation vs. 50 cents enclosed RR = 1.04 (95% CI, 0.87 to 1.25) 50 cents enclosed and $1 donation promised vs. 50 cents enclosed RR = 1.15 (95% CI, 0.97 to 1.36) $1 enclosed and $1 donation promised vs. 50 cents enclosed RR = 0.74 (95% CI, 0.60 to 0.91) $1 promised donation vs. $1 enclosed RR = 0.93 (95% CI, 0.79 to 1.10) 50 cents enclosed and $1 donation promised vs. $1 enclosed RR = 1.03 (95% CI, 0.88 to 1.19) $1enclosed and $1 donation promised vs. $1 enclosed RR = 1.27 (95% CI, 1.02 to 1.57) 50 cents enclosed and $1 donation promised vs. $1 promised donation RR = 1.39 (95% CI, 1.14 to 1.71) $1 enclosed and $1 donation promised vs. $1 promised donation RR = 1.10 (95% CI, 0.93 to 1.29) $1 enclosed and $1 donation promised vs. 50 cents enclosed and $1 donation promised
Hopkins and Podolak, 1983[207]	RCT	4	Non-health: views on emission control	Mechanics (USA)	Postal survey	***Incentive (enclosed) vs. no incentive:*** No incentive (286) $1 enclosed (95) (Also manipulated postal rates) See also *Table 19*	Primary: Response rate Secondary: Cost per response	No incentive 37% $1 enclosed 54%	RR = 1.45 (95% CI, 1.14 to 1.84) $1 enclosed vs. no incentive

continued

TABLE 25 contd *Incentives*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Primary and secondary outcomes	Response rates	RR (95% CI)
Paolillo and Lorenzi, 1984[251]	RCT	4	Non-health: topic unspecified	Business people (USA)	Postal survey	***Incentive (promised or enclosed) vs. no incentive:*** No incentive (100) Incentive (300) ***Promised vs. enclosed incentive:*** Promised (200) Enclosed (100) ***Nature of incentive:*** $1 enclosed (100) $2 promised (100) Promised entry in lottery for prizes of $50, $30 and $20 (100)	Primary: Response rates Secondary: None	No incentive 36% Incentive 46% Promised 37% Enclosed 65% $1 enclosed 65% $2 promised 41% Promised lottery entry 33%	***Incentive vs. no incentive:*** RR = 1.29 (95% CI, 0.96 to 1.72) any incentive vs. none RR = 1.81 (95% CI, 1.34 to 2.43) $1 enclosed vs. none RR = 1.14 (95% CI, 0.80 to 1.62) $2 promised vs. none RR = 0.92 (95% CI, 0.63 to 1.34) promised lottery vs. none ***Promised vs. enclosed incentive:*** RR = 1.76 (95% CI, 1.39 to 2.21) $1 enclosed vs. any promised RR = 1.59 (95% CI, 1.20 to 2.09) $1 enclosed vs. $2 promised RR = 1.97 (95% CI, 1.44 to 2.70) $1 enclosed vs. promised lottery ***Nature of (promised) incentive:*** RR = 1.24 (95% CI, 0.86 to 1.79) $2 promised vs. promised lottery
Blythe, 1986[255]	RCT	4	Non-health: use of clinical evaluation tools	Social work graduates (USA)	Postal survey	***Incentive (promised) vs. no incentive:*** No incentive (259) Promised entry in lottery for range of prizes (259)	Primary: Response rates Secondary: None	No incentive 40% Promised lottery entry 52%	RR = 1.34 (95% CI, 1.10 to 1.62) promised lottery vs. none
Trice, 1986[157]	RCT	4	Non-health: evaluation of hotel services	Hotel guests (USA)	Self-completion survey	***Incentive (promised) vs. no incentive:*** No incentive (1200) Promised reduction of $1 on hotel bill (1200) (Also manipulated: length of questionnaire; timing of administration (check-in or departure)) See also *Table 9*	Primary: Response rates Secondary: None	No incentive 16% Promised reduction 21%	RR = 1.34 (95% CI, 1.13 to 1.59) promised reduction vs. none
Biner, 1988[232]	RCT	4	Non-health: assessment of community needs	Householders with telephones (USA)	Postal survey	***Incentive (enclosed) vs. no incentive:*** No incentive (100) $1 enclosed (100) (Also manipulated covering letter (nature of appeal)) See also *Table 22*	Primary: Response rates Secondary: None	No incentive 34% $1 enclosed 68%	RR = 1.99 (95% CI, 1.44 to 2.75) $1 enclosed vs. none

continued

TABLE 25 contd *Incentives*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Primary and secondary outcomes	Response rates	RR (95% CI)
Dommeyer, 1988[257]	RCT	4	Non-health: product tampering and MCG scale	Householders with telephones (USA)	Postal survey	***Incentive (promised or enclosed) vs. no incentive:*** No incentive (100) Incentive (500) ***Promised vs. enclosed incentive:*** Promised (300) Enclosed (300) ***Nature of incentive:*** 25 cent coin enclosed (100) 25 cent cheque enclosed (100) 25 cent money order enclosed: 100 "Early-bird" incentive (100) $25 promised lottery (100)	Primary: Response rates Secondary: Speed of response Item non-response rates Mean score on MCG scale Cost of response	No incentive 37% Incentive 37.6% Promised 32% Enclosed 42% Coin enclosed 50% Cheque enclosed 37% Money order enclosed 38% "Early-bird" 33% Promised lottery entry 30%	***Incentive vs. no incentive:*** RR = 1.02 (95% CI, 0.77 to 1.34) any incentive vs. none RR = 1.35 (95% CI, 0.98 to 1.87) 25 cents coin vs. none RR = 1.00 (95% CI, 0.70 to 1.44) 25 cents cheque vs. none RR = 1.03 (95% CI, 0.72 to 1.47) 25 cents money order vs. none RR = 0.89 (95% CI, 0.61 to 1.30) "early bird" vs. none RR = 0.81 (95% CI, 0.55 to 1.20) promised lottery vs. none ***Promised vs. enclosed incentive:*** RR = 1.32 (95% CI, 1.04 to 1.69) enclosed vs. promised ***Nature of incentive:*** RR = 0.74 (95% CI, 0.54 to 1.02) cheque vs. coin RR = 0.76 (95% CI, 0.55 to 1.04) money order vs. coin RR = 0.66 (95% CI, 0.47 to 0.93) "early bird" vs. coin RR = 0.60 (95% CI, 0.42 to 0.86) promised lottery vs. coin RR = 1.03 (95% CI, 0.72 to 1.47) money order vs. cheque RR = 0.89 (95% CI, 0.61 to 1.30) "early bird" vs. cheque RR = 0.81 (95% CI, 0.55 to 1.20) promised lottery vs. cheque RR = 0.87 (95% CI, 0.60 to 1.26) "early bird" vs. money order RR = 0.79 (95% CI, 0.53 to 1.17) promised lottery vs. money order RR = 0.91 (95% CI, 0.60 to 1.37) promised lottery vs. "early bird"

continued

TABLE 25 contd *Incentives*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Primary and secondary outcomes	Response rates	RR (95% CI)
Hubbard and Little, 1988[258]	RCT	4	Non-health: satisfaction with banking and financial services	Householders with telephones (USA)	Postal survey	***Incentive vs. no incentive:*** No incentive (400) Incentive (1600) ***Promised vs. enclosed incentive:*** Promised (800) Enclosed (800) ***Nature of incentive:*** Promised donation of $1 (400) 25 cents enclosed (400) $1 enclosed (400) Promised entry in lottery with prize of $200 (400)	Primary: Response rates Secondary: Speed of response Cost of response Quality of response	No incentive 41% Incentive 53% Promised 43% Enclosed 62% Promised donation of $1 34% 25 cents enclosed 57% $1 enclosed 68% Promised entry in lottery with prize of $200 52%	Incentive vs. no incentive: RR = 1.30 (95% CI, 1.14 to 1.47) any incentive vs. none RR = 0.83 (95% CI, 0.69 to 0.99) promised donation vs. none RR = 1.40 (95% CI, 1.21 to 1.62) 25 cents enclosed vs. none RR = 1.68 (95% CI, 1.46 to 1.92) $1 enclosed vs. none RR = 1.28 (95% CI, 1.10 to 1.49) promised lottery vs. none Promised vs. enclosed incentive: RR = 1.46 (95% CI, 1.33 to 1.61) enclosed vs. promised Nature of incentive: RR = 1.69 (95% CI, 1.44 to 1.99) 25 cents vs. promised donation RR = 2.03 (95% CI, 1.74 to 2.37) $1 vs. promised donation RR = 1.54 (95% CI, 1.31 to 1.83) promised lottery vs. promised donation RR = 1.20 (95% CI, 1.07 to 1.34) $1 vs. 25 cents RR = 0.91 (95% CI, 0.80 to 1.04) promised lottery vs. 25 cents RR = 0.76 (95% CI, 0.68 to 0.85) promised lottery vs. $1
Suhre, 1989[163]	RCT	3	Non-health: educational innovation and effectiveness	School principals (The Netherlands)	Postal survey	***Incentive (promised) vs. no incentive:*** No incentive (332) Incentive (663) ***Nature of incentive:*** Promised $10 incentive (332) Promised entry in lottery with prizes worth total of $700 (331) (Also manipulated pagination) See also *Table 10*	Primary: Response rates Secondary: None	Insufficient data reported to calculate response rates	Reported that response rates were similar (overall response rate 52%) but insufficient data presented to calculate RR

continued

TABLE 25 contd *Incentives*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Primary and secondary outcomes	Response rates	RR (95% CI)
Biner and Barton, 1990[234]	RCT	4	Non-health: assessment of community needs	Householders with telephones (USA)	Postal survey	***Size of incentive:*** 25 cents enclosed (100) $1 enclosed (100) (Also manipulated covering letter (portrayal of incentive)) See also *Table 22*	Primary: Response rates Secondary: None	25 cents enclosed 50% $1 enclosed 69%	RR = 1.41 (95% CI, 1.10 to 1.80) $1 vs. 25 cents
Brennan, 1992[259]	RCT	4	Non-health: Study 1: Direct marketing of financial services Study 2: Sponsorship Study 3: Use and views of fruit juice Study 4: Catalogue vs. retail shopping Study 5: Personal finances	Study 1: Client's financial services "hot property" list Studies 2–5: People on electoral register (New Zealand)	Postal survey	***Incentive (enclosed) vs. no incentive:*** No incentive: Study 1 (200) Study 2 (250) Study 3 (200) Study 4 (192) Study 5 (100) 50 cents enclosed: Study 1 (200) Study 2 (250) Study 3 (200) Study 4 (192) Study 5 (100)	Primary: Response rates Secondary: Speed of response Cost per response	Study 1: No incentive 72% 50 cents enclosed 81% Study 2: No incentive 54% 50 cents enclosed 66% Study 3: No incentive 56% 50 cents enclosed 64% Study 4: No incentive 64% 50 cents enclosed 70% Study 5: No incentive 58% 50 cents enclosed 75%	Study 1: RR = 1.13 (95% CI, 1.01 to 1.26) 50 cents enclosed vs. none Study 2: RR = 1.22 (95% CI, 1.06 to 1.41) 50 cents enclosed vs. none Study 3: RR = 1.14 (95% CI, 0.97 to 1.34) 50 cents enclosed vs. none Study 4: RR = 1.09 (95% CI, 0.95 to 1.25) 50 cents enclosed vs. none Study 5: RR = 1.29 (95% CI, 1.06 to 1.58) 50 cents enclosed vs. none

continued

TABLE 25 contd *Incentives*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Primary and secondary outcomes	Response rates	RR (95% CI)
Gajraj *et al.*, 1990[183]	Non-random concurrent controlled study	4	Non-health: issues of energy conservation	Customers of major public utility (Canada)	Postal survey	***Incentive vs. no incentive:*** No incentive (100) Incentive (600) ***Promised vs. enclosed incentive:*** Promised (300) Enclosed (300) ***Nature of incentive:*** 50 cents enclosed (100) 50 cents promised (100) Pen enclosed (100) Pen promised (100) Definite inclusion in share of potential winnings from public lottery (100) Promised inclusion in share of potential winnings from public lottery (100) Lottery incentives comprised 5 lottery tickets at $10.00 each	Primary: Response rates Secondary: Speed of response Cost per response	No incentive 34% Incentive 48% Promised 43% Enclosed 53% 50 cents enclosed 62% 50 cents promised 42% Pen enclosed 41% Pen promised 43% Definite inclusion in share of potential winnings from public lottery 55% Promised inclusion in share of potential winnings from public lottery 45%	***Incentive vs. no incentive:*** RR = 1.41 (95% CI, 1.06 to 1.88) any incentive vs. none RR = 1.82 (95% CI, 1.33 to 2.49) 50 cents enclosed vs. none RR = 1.24 (95% CI, 0.86 to 1.77) 50 cents promised vs. none RR = 1.21 (95% CI, 0.84 to 1.73) pen enclosed vs. none RR = 1.26 (95% CI, 0.89 to 1.80) pen promised vs. none RR = 1.62 (95% CI, 1.17 to 2.24) lottery enclosed vs. none RR = 1.32 (95% CI, 0.93 to 1.88) lottery promised vs. none ***Promised vs. enclosed incentive:*** RR = 1.22 (95% CI, 1.03 to 1.44) any enclosed vs. any promised RR = 1.48 (95% CI, 1.12 to 1.95) 50 cents enclosed vs. 50 cents promised RR = 0.95 (95% CI, 0.69 to 1.32) pen enclosed vs. pen promised RR = 1.22 (95% CI, 0.92 to 1.62) lottery enclosed vs. lottery promised ***Nature of incentive (controlling for whether promised or enclosed):*** RR = 0.66 (95% CI, 0.24 to 0.75) pen vs. 50 cents (enclosed) RR = 0.89 (95% CI, 0.43 to 1.32) lottery vs. 50 cents (enclosed) RR = 1.34 (95% CI, 1.00 to 3.08) lottery vs. pen (enclosed) RR = 1.02 (95% CI, 0.74 ,1.41) pen vs. 50 cents (promised) RR = 1.07 (95% CI, 0.78 to 1.47) lottery vs. 50 cents (promised) RR = 1.05 (95% CI, 0.77 to 1.43) lottery vs. pen (promised)

MCG, Morality–Conscience–Guilt

TABLE 26 *Feedback of results*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Primary and secondary outcomes	Response rates	RR (95% CI)
Powers and Alderman, 1982[154]	RCT	4	Non-health: reactions to scholastic tests and related materials	School students (USA)	Postal survey	***Offer vs. no offer:*** No offer of results (1003) Offer of results (1004) (Also manipulated length of questionnaire) See also *Table 9*	Primary: Response rates Secondary: None	No offer of results 46% Offer of results 50%	RR = 1.10 (95% CI, 1.01 to 1.21) offer vs. no offer
Dommeyer, 1985[240]	RCT	4	Non-health: Mind Inventory Catalogue and Tax Survey (one or other of these scales)	Business students (USA)	Postal survey	***Offer vs. no offer:*** No offer of results (210) Offer of results summary (210) (Also manipulated saliency) See also *Table 24*	Primary: Response rates Secondary: Non-response bias Speed of response Item non-response rates Request for results	No offer of results 43% Offer of results 42%	RR = 0.98 (95% CI, 0.78 to 1.22) offer vs. no offer
Dommeyer, 1989[262]	RCT	4	Non-health: computer counterfeiting	PC owners, manufacturers and retailers (USA)	Postal survey	***Offer vs. no offer:*** No offer of results (300) Offer of results (600) ***How results offered:*** Offered in covering letter (300) Offered in "lift" letter (300)	Primary: Response rates Secondary: Speed of response Request for results Item non-response rates Cost per response	No offer of results 24% Offer of results 20% Offer in covering letter 23% Offer in "lift" letter 16%	***Offer vs. no offer:*** RR = 0.81 (95% CI, 0.63 to 1.05) any offer vs. no offer RR = 0.96 (95% CI, 0.72 to 1.28) offer in cover letter vs. no offer RR = 0.67 (95% CI, 0.48 to 0.93) offer in "lift" letter vs. no offer ***How results offered:*** RR = 0.75 (95% CI, 0.41 to 1.36) "lift" letter vs. cover letter
Green and Kvidahl, 1989[227]	RCT	3	Non-health: opinions on application of research findings to teaching	School teachers (USA)	Postal survey	***Offer vs. no offer:*** No offer of results (≅300) Results offered (≅300) (Also manipulated personalisation) See also *Table 21*	Primary: Response rates Secondary: None	No significant difference in overall response rates (actual data not reported)	Insufficient data to calculate RR

TABLE 27 *Miscellaneous*

Reference	Study design	Quality score	Topic	Respondents (country)	Mode of admin.	Factors manipulated (sample size)	Primary and secondary outcomes	Response rates	RR (95% CI)
Elkind *et al.*, 1986[209]	RCT	4	Health: experience of abuse and harassment by patients	Professional psychologists (USA)	Postal survey	***Addressing of return envelope:*** Envelope with return address added by rubber stamp (250) Preprinted return address (250) (Also manipulated postage type) See also *Table 19*	Primary: Response rates Secondary: None	Rubber stamp 63% Preprinted 66%	RR = 1.04 (95% CI, 0.92 to 1.19) preprinted vs. rubber stamp
Salvesen and Vatten, 1992[264]	RCT	3.5	Health: harmful effects of screening	Women participating in RCT on ultrasound (Norway)	Postal survey	***Enclosure of relevant newspaper article:*** No article enclosed (358) Article enclosed (358)	Primary: Response rates Secondary: None	Data presented as a survival curve; not possible to calculate response rates directly but women who did not receive the newspaper article were less likely to respond	Reported odds ratio = 0.79 (95% CI, 0.66 to 0.96)
Lovelock *et al.*, 1976[263]	Cross-sectional study (included because of novel approach)	Not graded for quality	Non-health: transportation attitudes and behaviour	Householders (USA)	Self-completion survey	***Delivery of questionnaire by hand:*** Total sample size 1465 Eligible households 1361 Net sample 1183 Total questionnaires distributed (aimed to distribute 2 per household where possible) 1851	Primary: Response rates (return of completed questionnaire; return of usable questionnaire) Secondary: Contact rates % accepting questionnaire % of households responding	Questionnaires returned completed 1455/1851 (79%) Usable questionnaires returned 1407/1851 (76%)	Not applicable

Chapter 7

Summary of conclusions

Throughout, care is required in extrapolating findings from one setting to another (e.g. market research to health-related surveys) and from one culture to another (e.g. the USA to the UK).

Mode of administration (chapter 3)

Telephone interviews versus postal surveys

- Telephone interviews generally obtain higher response rates than postal surveys.
- Evidence from a single study suggested that rates of item non-response may be higher for postal surveys.
- There is little consensus about the benefits of telephone and postal surveys on parameters of non-response bias, quality of response, anonymity or cost.

Face-to-face interviews versus self-completion questionnaires

- Face-to-face interviews tend to yield higher response rates.
- Evidence from a single study suggests that respondents may be more likely to give no answer at all or say "don't know" in an interview than in a self-completion questionnaire or card sort.
- There is a lack of unequivocal evidence to support the view that postal survey participants respond more truthfully to sensitive issues or make more critical or less socially acceptable responses than when face-to-face with an interviewer.

Telephone versus face-to-face interviews

- Telephone interviews may be quicker than face-to-face interviews.
- There is no consistent evidence of the relative superiority of face-to-face interviews or telephone interviews on the parameters of instrument response rate, eliciting sensitive information and item non-response rates.

Computer-assisted versus paper-based self-completion questionnaires

- Findings on the effects of computer-assisted versus paper-based questionnaires on response rates and response quality are equivocal.
- Evidence from a single study suggests that respondents to computerised questionnaires may use a wider range on rating scales.
- Quicker responses may be obtained using computer-assisted questionnaires.
- There is no clear evidence that responses to sensitive questions differ between computer-assisted and paper-based modes.

Computer-assisted telephone interviewing versus conventional telephone interviewing

- Interviewer variability may be lower with CATI.

General

- No one mode of administration consistently outperforms all others.

Question wording and sequencing (chapter 4)

Some caution is required in extrapolating the findings of previous research to health surveys. Few of the identified studies were health related; generalisability of findings to this field may be limited. For many of the topics investigated only one or two relevant studies were identified that met the review quality criteria. Moreover, theories of both cognition and response formulation, as well as empirical evidence, suggest that the effects of question and response category wording and ordering may vary with the mode of administration, again indicating a need for caution in interpreting and applying findings from interview surveys to self-completion questionnaires or vice versa.

Question wording

- Question wording and format can influence both whether or not an opinion is given and what opinion is given.
- Open-ended questions produce more non-common category responses than closed questions, but most additional categories are small and miscellaneous. The use of either question form will ordinarily lead to similar conclusions. However, expert opinion suggests that open-ended questions still remain important in development stages and pilot studies; using open-ended questions at these

stages allows researchers to generate appropriate response categories for closed questions.
- Survey participants employ a wide range of cognitive processes in formulating responses to behavioural frequency questions, including episodic enumeration (i.e. recalling and counting specific instances on which the behaviour occurred) and rate processing (i.e. aggregating from the "normal" rate at which the behaviour takes place in a unit of time, such as a week). Task conditions, such as the time-framing of a question, will influence the processes employed.
- The wording of filter questions asking whether the respondent has knowledge of or has thought about an issue can significantly affect the percentages of "don't know" responses elicited at a subsequent substantive question, particularly for topics that are less familiar to the informant. Conversely, the content of the question can have an important independent effect on "don't know" responses, regardless of the filter wording. The use of filter questions can alter the conclusions drawn.
- Giving a second substantive choice on attitudinal questions increases the likelihood of respondents expressing an opinion.
- Acquiescence effects tend to be negatively related to the educational level of the respondent.
- Response bias may be introduced by the use of mixed grammar chains (e.g. elliptical versus non-elliptical structures). Elliptical structure questions (those in which the verb is omitted) produce both more agreement and more disagreement while non-elliptical ones produce more neutral responses.
- Although the inclusion of negatively phrased items may theoretically control or offset acquiescence tendencies, their actual effect may be to reduce response validity.
- The interpretation of questions that include prestige names is complicated by the fact that participants respond not only on the basis of the content of issues but also on the basis of the names. Prestige names represent both additional stimuli and additional sources of variance to be explained.

Question sequencing

- Question order effects may influence overall response rates and increase sample composition bias in a range of ways; the direction and strength of these effects can vary with topic, context and study population.
- Questionnaire design, particularly the apparent relevance of opening items, may influence people's motivation to complete the instrument; the more salient and relevant these items are, the greater the likelihood of response.
- The received wisdom that questions should be grouped by topic and ordered so that related topics are adjacent to one another ignores the issue of context effects; yet research suggests that such effects are common. Researchers need to balance the risk of context effects with the desirability of coherence and continuity.
- Topic ordering within a questionnaire may differentially affect response rates among different attitudinal groups.
- Question order effects may not be ubiquitous, but evidence suggests that general questions should precede specific questions.
- Context effects may bias estimates of the prevalence of attitudes and behaviour. If otherwise identical questions are posed in a different order or a different context across questionnaires (either at the same point in time or in longitudinal studies), apparent differences in response patterns may reflect context effects rather than true differences between respondents.
- Context effects are especially likely when researchers attempt to summarise complex issues in a single general item.
- Context effects may be larger when respondents' beliefs about a target issue are both mixed and important to them.
- Prior items in a questionnaire may exert a "carry-over" effect by priming respondents about their beliefs/attitudes towards a particular topic.
- "Buffering" of items may reduce context effects, but is unlikely to eliminate them completely.
- Question order effects tend to be consistent across gender and educational levels and so are as much of a concern in surveys on restricted populations as in those on general populations.
- Context effects may be lessened but not entirely eliminated in self-completion questionnaires.
- Evidence suggests that scores on disease-specific and generic health status measures in health outcomes questionnaires are unaffected by their position relative to one another. This question was investigated in only one disease area with a relatively bounded impact on overall health status, so further studies are required to determine whether the results are generalisable.

Response format

- The inclusion of middle position, no opinion and "don't know" response options seems generally preferable for attitudinal questions, although this may be less important for factual questions.
- Providing informants with an opportunity to have no opinion may avoid spurious representativeness.
- The "middle response" category does not necessarily represent a position of neutrality and its exclusion may produce invalid results.

- The wording of response categories is as critical as that of question wording because ambiguity in their meaning contributes to response order effects.
- The order of response alternatives may affect both the distribution of responses to individual items and the associations between these and other items.
- Recency effects (a tendency to choose the last response option) appear to be uncommon in self-completion questionnaires.
- Findings on the relative merits of single-step and two-step approaches to presenting response categories for attitudinal questions in telephone surveys are equivocal.
- Evidence about the labelling of response categories is inconsistent, but fully defining scales may act as a check on leniency errors.
- A remote scale format in which the response categories are at a distance from the question appears to be associated with a tendency towards neutrality of response.
- The inclusion of a space for free comment may increase response rates.

Questionnaire appearance (chapter 5)

Few of the identified studies were health related; generalisability of findings to this field may therefore be limited. For many of the topics investigated, only one or two relevant studies that met the quality criteria were identified. This means that caution should be exercised in interpreting and extrapolating findings. Nevertheless, in the absence of empirical evidence, much expert opinion makes sound sense and is supported by theories of cognition, perception and pattern recognition,[145] but it must be noted that some of the recommendations made in the key texts are based on "old technology" (in particular, typewritten documents) and fail to take into account the facilities afforded by modern word-processing and desktop publishing software, by current reprographics facilities, or by the use of scannable (OMR and OCR) questionnaires (see chapter 3).

Length of questionnaire

- Findings with respect to the impact of questionnaire length on response rates are equivocal.
- A saliency by questionnaire length interaction has been demonstrated in previous reviews; questionnaires on highly salient (relevant or interesting) topics (as will be the case in many surveys of patients or health professionals) can probably be longer than questionnaires on more general topics or those for a general population.
- There is the potential for response bias, due to fatigue or carelessness, in the latter part of long questionnaires, particularly with respect to answers to "item sets".

Pagination

- In terms of response rates, the superiority of a booklet format questionnaire, or of an A4 (as opposed to A5 or other size) format has not been demonstrated.
- Findings with respect to the provision of "white space" are equivocal.

Paper colour

- Questionnaire colour has not been shown to have a significant impact on response rates.

Question and response category formats

- Differences in response rates or response quality between a "tick the box" format and a "circle the number" format have not been shown to be significant.

Enhancing response rates (chapter 6)

Caution is required in interpreting results, particularly in comparing findings across studies, because of the heterogeneity of study populations, survey topics and factors manipulated.

Timing of survey

- Response rates do not appear to be affected by the day of posting.
- The month of posting may affect response rates, but this may be topic specific.

Number and relative timing of contacts

- Response rates can be increased through multiple contacts.
- Although both prenotification and follow-up contacts are effective in stimulating response rates, the latter is likely to be more powerful.

Prenotification contacts

- Prenotification is effective in increasing response rates.
- Prenotification by letter may be more effective than prenotification by telephone.
- High involvement methods of prenotification (e.g. "foot-in-door" approaches) have not been shown conclusively to improve response rates over simple prenotification; such high involvement approaches are really feasible only when

telephone or personal approaches to prenotification are made.

Follow-up contacts (reminders)

- Follow-up contacts are highly effective in increasing response rates.
- There is no conclusive evidence that a "threat" of further follow-ups made in a reminder letter enhances response rates.
- Including a duplicate questionnaire with the first reminder does not seem to have a significant impact on response rates, but the inclusion of a replacement questionnaire with a second reminder seems to be effective.
- There is no conclusive evidence that special mailing techniques for final reminders are superior to standard mailing. Postcard reminders appear to be as effective as letters and are generally cheaper (although there may be concerns of confidentiality in health surveys).

Postage rates and types

- Findings from both primary studies and previous reviews show no consistent advantage of class of mail, or of stamped envelopes over reply-paid envelopes.

Confidentiality/anonymity

- Assurances of complete anonymity do not significantly improve response rates and may indeed have a detrimental effect.

Personalisation

- There is little conclusive evidence of the advantages of personalisation of covering letters and envelopes *per se*, but personalisation may interact with such factors as the nature of the appeal made in the covering letter and assurances of confidentiality.

Covering letters

- Traditional-style letters are more effective than novel approaches.
- There is little conclusive evidence that the characteristics of the signatory affect response rates.
- Response rates do not appear to be positively related to handwritten signatures or colour of ink.
- No one type of appeal in the covering letter offers a consistent advantage; rather, the nature of the appeal should be matched to the anticipated motivations of the recipients.

Time cues and deadlines

- A short time cue can be effective in stimulating responses.
- Specification of a deadline for responding may increase the speed of response (and thereby reduce the number of reminders needed), but it may have no effect on overall response rates.

Sponsorship

- The impact of sponsorship appears to be situation and location specific.

Saliency

- A salient (interesting, relevant and current) topic is effective in enhancing response rates.

Incentives

- Incentives are generally an effective means of increasing response.
- Financial incentives are likely to be more effective than non-monetary incentives of similar value.
- Enclosed incentives are more effective than promised incentives.

Feedback of results

- Offering feedback of survey results is generally not effective in stimulating response.

Miscellaneous

- Personal delivery of questionnaires for self-completion may offer some advantages but the cost-effectiveness of this approach may be situation specific.

Chapter 8

Summary of recommendations for practice

In each substantive section below, recommendations made on the basis of evidence from one or more high-quality primary studies are separated from those based on expert opinion, previous literature reviews, theories of respondent behaviour, and the accumulated experience of the review team with respect to the conduct of surveys. When findings from primary research studies were negative or equivocal, the authors indicate that their recommendations are derived from these findings, rather than being directly based upon them. Where evidence from primary studies is reinforced by expert opinion or previous literature reviews, or is underpinned by theory, they highlight this fact.

Mode of administration (chapter 3)

Findings from high-grade primary studies were equivocal, suggesting that no single mode of administration is superior in all respects or in all settings. The choice of mode of administration should therefore be made on a survey-by-survey basis, taking into account:

- study population
- survey topic
- sampling frame availability and quality
- sampling method
- volume of data to be collected
- complexity of data to be collected
- resources available.

Before embarking on a survey in a particular setting, with a particular population, or on a particular topic, the researcher should review the literature carefully to ascertain the appropriateness of the survey method in general, and of different modes of survey administration in particular (including the likely impact of interviewer characteristics), in those particular circumstances.

Question wording and sequencing (chapter 4)

Some caution is required in extrapolating the findings of previous research to health surveys. Few of the identified studies were health related; generalisability of findings to this field may be limited. Moreover, the majority of the identified studies were carried out in the USA and generalisability to the UK may be limited by linguistic differences. For many of the topics investigated, only one or two relevant studies that met the quality criteria were identified; therefore, even evidence-based recommendations are founded on limited findings from previous research.

Recommendations with an evidence base from one or more high-grade primary comparative studies

Question wording

- Efforts to increase response accuracy should take into account the range of cognitive processes involved in response formulation and the potential impact of task variables such as the likely salience and temporal regularity of events, the method of survey administration, and question design issues such as the time-frame. (Recommendation based on evidence from primary research studies and on theories of response formulation.)
- Open-ended questions should be used sparingly, particularly in self-completion questionnaires; however, careful piloting and pretesting using open-ended questions should be carried out to ensure that the response categories presented in closed questions adequately represent the likely range of responses. (Recommendation based on evidence from primary research studies and on expert opinion.)
- Combining elliptical and non-elliptical structure questions can bias results, so this should be avoided where possible.
- Until further investigations have been carried out and firmer evidence is available, caution should be exercised in the use of negatively phrased attitudinal items.
- The implications of including or excluding filter questions on response distributions should be considered.
- Researchers should be aware of the difficulties inherent in interpreting responses to survey questions involving prestige names and avoid their use wherever possible.

Question sequencing

- Researchers should be aware of the potential for question order effects in self-completion

questionnaires as well as in interview surveys, and follow expert opinion about questionnaire design accordingly.
- General questions should precede specific questions. (Recommendation based on evidence from primary research studies and expert opinion.)
- There is evidence that "buffering" of questions is unlikely to eliminate context effects, so researchers should adhere to the common survey practice of blocking questions by topic. (Recommendation based on evidence from primary research studies and expert opinion.)
- Where there is evidence that respondents may have stronger opinions on some survey topics than on others, the priority of their concerns should be determined and the survey instrument assembled to reflect them.
- Demographic questions should be placed at the end of the questionnaire. (Recommendation based on limited evidence from primary research studies combined with expert opinion.)
- Given the current lack of evidence of any ordering effects, ordering of generic and disease-specific measures should follow the rules for general versus specific questions. (Recommendation based on limited evidence from primary research studies combined with expert opinion.)

Response format
- The middle response category for attitude/opinion questions does not necessarily represent a position of neutrality, so it should be included.
- For factual questions, the "don't know" response may reasonably be omitted.
- If a remote scale format is used in self-completion questionnaires, the stem question should be repeated every three or four questions.
- An open space for free comment should be included in self-completion questionnaires.

Recommendations derived solely from theories of cognition and response formulation and/or expert opinion

Careful piloting of questions and their associated response categories is strongly advised, particularly when the questions have been developed especially for that survey, or when questions or scales used in a different setting or with a different population are to be used. Context gives meaning to questions, and question ordering effects are rife, so questions should be piloted in context rather than in isolation. Cognitive interviewing techniques[134–137] are useful in gaining an understanding of how respondents understand and interpret questions and the thought processes (e.g. episodic enumeration) and heuristics (e.g. generalising from the most recently retrieved memory) they employ in responding (see appendix 1).

Question wording
- General principles of questionnaire wording (*Box 2*; see p. 60) should be maintained.
- Note that the question stem and associated response categories combine to convey meaning; one should not be designed in isolation from the other.

Question sequencing
- In situations where investigators are uncertain about the impact of question order on results, the order should be randomised.
- In longitudinal studies, or in those being carried out in multiple settings, the same question ordering should be maintained over time and across locations.

Response formats
- Response categories for closed questions should be mutually exclusive (i.e. unambiguous, not overlapping) and collectively exhaustive (all contingencies catered for, if necessary by the inclusion of an option of "Other, please specify").
- It must be noted that the nature of the response categories gives a subtle message about the range of ideas/concepts that the respondent should be thinking about.
- The evidence is inconsistent, so it may be preferable to label all response categories rather than only the end-points.

Questionnaire appearance (chapter 5)

Few of the identified studies were health related; generalisability of findings to this field may therefore be limited. For many of the topics investigated, only one or two relevant studies that met the quality criteria were identified. This means that caution should be exercised in interpreting and extrapolating findings.

Although no literature was identified on the topic, computer-scannable questionnaires (OCR and OMR) are likely to assume greater importance in the future. Design principles for scannable questionnaires should be guided by the hardware and software to be used.

Recommendations with an evidence base from one or more high-grade primary comparative studies

Length of questionnaire
- Avoid excessively long questionnaires, especially if the topic is likely to be of low saliency to the respondents. (Recommendation based on the

evidence base from primary studies and from previous reviews.)
- Avoid crowding questions or reducing "white space" in a desire to reduce apparent length. (Recommendation based on limited evidence from primary studies and on expert opinion.)

Response formats
- Use a "circle the number" format rather than a "tick the box" format in self-completion questionnaires. (Recommendation based on limited evidence from one primary study and on expert opinion.)

Recommendations based on theories of perception and cognition and/or expert opinion

General
- In surveys of patients, it is likely that a significant proportion of the target sample will have some degree of visual impairment. The needs of such individuals should be taken into account in designing self-completion questionnaires.

Pagination
- Use a booklet format with double-sided printing.
- Use standard-sized paper (A4 folded to A5 booklet or A3 folded to A4 booklet, as dictated by the length of the questionnaire).

Placement of questions within pages
- Avoid splitting a question, its associated response categories and instructions for answering over two pages.
- In questions where the list of response categories is too long to fit on a single page, continue the response categories on a facing page if possible; otherwise repeat the question on the subsequent page.
- Do not ask respondents to do two things (e.g. rating and ranking) in responding to one question.
- When one question is logically dependent upon another, make every effort to place both on the same page.
- Avoid placing a short question at the foot of a page, especially if preceded by a long question with a number of subparts.

Use of "white space"
- Leave sufficient space for responses to open-ended questions.
- Do not use lines for responses to open-ended questions, unless only a short response (i.e. a number or a few words) is required.

Paper colour
- Paper colour has not been shown to have a significant impact on response rates (evidence base), so choose white paper or a light tint to enhance legibility.
- Consider the use of coloured covers to distinguish questionnaires.

Print details
- Use a font size of at least 10 points; a larger font size (up to 14 or 16 points, depending on type face) is desirable if it is anticipated that respondents may have some visual impairment, for example in surveys of older people.
- Use a distinct typeface and avoid excessive use of italics and upper case characters, especially in self-completion questionnaires.

Cover design
- The front cover of the questionnaire should contain the title of the survey (not "Questionnaire on X topic..."), the identity of the organisation carrying it out and, for self-completion surveys, a neutral graphic illustration.
- The back cover of the questionnaire should provide some blank space for respondents' open comments, and should specify the address of the organisation conducting the survey (if not on the front cover) and say "thank you" to the respondent.

Question and response category format
- Use elements of brightness, colour, shape and location to "steer" the respondent through the questionnaire.
- Maintain a consistent format throughout the questionnaire.
- Use a vertical response format (*Figure 1*; see p. 93) for closed-ended questions, except for rating scales.
- Use a horizontal response format (*Figure 2*; see p. 93) for item sets involving the same response categories throughout, and in rating scales.
- Consider natural reading style (i.e. left to right, and horizontally orientated) in placing headings and codes for responses.
- Use graphical means (e.g. arrows and boxes) to indicate skip patterns.
- Place instructions and directions at the point where they are required; if a series of questions involves turning a page, it may be necessary to repeat instructions on the new page.

Enhancing response rates (chapter 6)

Relatively few of the studies were health related; the generalisability of findings to this field (where

response rates are typically better in any case) may be limited. It should also be noted that, for many of the reviewed areas, the findings from both primary studies and from previous reviews were at best equivocal, and, in some cases, contrary to expert opinion, as set out in key texts on survey design.[1,5,16,40,194] In the recommendations that follow, those derived from mixed or negative findings in previous research are highlighted.

Despite mixed findings, what is apparent is that there is no single method of enhancing response rates that is applicable in all settings. Instead, the choice of techniques should be informed by consideration of the likely barriers and motivational factors for each particular survey topic and study population. The frameworks presented by Dillman[1] and by Brown and colleagues[140] form a useful basis for deliberation.

In assessing potential methods, the researcher should consider not only the likely impact on response rates but also the potential for non-response and sample composition biases, response bias and item non-response effects, as well as the implications for resources of time, money, personnel and materials. The marginal benefits of intensive approaches to enhancing response rates may be outweighed by the marginal costs.

Manipulation of a single factor is unlikely to prove fruitful. Instead, the researcher should consider the total "package" of questionnaire wording, questionnaire appearance, general motivational factors (anonymity/confidentiality; personalisation; nature of appeal; other aspects of the covering letter; sponsorship; saliency), mechanical and perceptual factors (timing of survey; number, timing and method of contacts; postage rates and types), and financial and other incentives.[1]

Recommendations with an evidence base from one or more high-grade primary comparative studies

General

- Consider the possibility of interactions between factors and take care to avoid apparent mismatches (e.g. a highly personalised letter combined with an assurance of total anonymity). (Recommendation based on observed interaction effects in primary studies and on expert opinion.)

Number and relative timing of contacts

- Use multiple contacts (at least one contact in addition to the initial mailing of the questionnaire). Note, however, that ethics committees may consider highly intensive contact procedures (more then three contacts) to be overly intrusive (Key L, Newcastle and North Tyneside Joint Research Ethics Committee, Newcastle upon Tyne: personal communication, 1999). (Recommendation based on evidence from primary studies, findings from previous reviews and expert opinion.)
- Consider both prenotification and follow-up contacts.
- If resources are limited, concentrate on follow-up contacts rather than prenotification. (Recommendation based on evidence from primary studies, findings from previous reviews and expert opinion.)

Prenotification

- Consider prenotification, preferably by letter, to alert target respondents to the arrival of the questionnaire. (Recommendation based on evidence from primary studies, findings from previous reviews and expert opinion.)

Follow-up (reminders)

- Use at least one reminder to non-respondents. (Recommendation based on evidence from primary studies, findings from previous reviews and expert opinion.)
- Match the appeal in the reminder letter to the perceived motivations of the study population; a consensual approach may be appropriate to some groups while a "threat" of further follow-up may be more effective with others. (Recommendation derived from lack of consistent findings from primary studies and previous reviews; recommendation is supported by theories of respondent behaviour.)
- If initial non-response is perceived to be related to non-delivery or mislaying of the questionnaire, consider including a duplicate questionnaire with the reminder. If two or more reminders are being used, it may be appropriate to wait until the second or subsequent reminder to enclose the duplicate questionnaire. (Recommendation derived from lack of consistent findings from primary studies and previous reviews; this recommendation is supported by expert opinion.)
- Choose a mode of contact for reminders that is appropriate to the survey topic and study population. Intensive techniques such as certified or recorded delivery mailing may be considered by target respondents or by ethics committees to be overly intrusive or unduly coercive (Key L, Newcastle and North Tyneside Joint Research Ethics Committee, Newcastle upon Tyne: personal communication, 1999). Although postcard reminders may be cost-

effective, concerns regarding confidentiality may preclude their use in surveys on health-related topics. (Recommendation derived from lack of consistent findings from primary studies and previous reviews; this recommendation is supported by expert opinion.)

Anonymity/confidentiality

- In general, total anonymity is not appropriate. Use coded (i.e. numbered and therefore identifiable) questionnaires to facilitate follow-up and record linkage. It is appropriate to be explicit in a covering letter or information sheet about how the code number will be used (i.e. to keep a check on who has responded and thereby to allow non-respondents to be followed up). (Recommendation based on evidence from primary studies and expert opinion.)

Postage rates

- For convenience, use franking rather than postage stamps for outgoing mail and use business reply envelopes for return of questionnaires. The choice between first and second class mail should involve consideration of the relative costs, the speed with which results are required, and whether it is anticipated that respondents will be aware of or influenced by the class of mail. (Recommendation derived from lack of consistent findings from primary studies and previous reviews; this recommendation is supported by the accumulated experience of the review team.)

Personalisation

- In surveys of general populations, personalisation may offer no significant advantage. However, personalisation of covering letters is likely to be appropriate if the message in the letter suggests personal knowledge of the circumstances of the recipient or uses a self-interest appeal. For example, a personalised approach may be appropriate in a survey of patients selected because they have a particular health problem. Personalisation may also be appropriate when the target respondents are in fact personally or professionally known to the sender. (Recommendation derived from lack of consistent findings from primary studies and previous reviews.)

Covering letter: style and content

- Use a traditional letter format, including headed notepaper. (Recommendation based on evidence from one primary study and on expert opinion.)
- In most circumstances, a facsimile signature is likely to be adequate. However, care should be taken to match the degree of personalisation of the signature to personalisation of the body of the letter.
- No single type of covering letter appeal is universally appropriate. Rather, the nature of the appeal made in the covering letter should be based on the perceived motivations of the study population and should be ethically sound. (Recommendation derived from lack of evidence from primary studies of any consistent advantage of one particular type of appeal; this recommendation is supported by theories of respondent behaviour.)
- Consider including in the covering letter a realistic indication of the time required for completion of the questionnaire.
- Consider specifying a deadline for response, especially if a timely response is of the essence.

Sponsorship

- If ethical and practical constraints permit, choose a study sponsor appropriate to the survey topic and study population; manipulate the covering letter and return address appropriately. In surveys on health-related topics, response rates may be enhanced if the covering letter purports to come from the recipients' health-care provider. However, consideration should be given to whether this approach may induce response bias (for example, if patients believe their doctor is going to see their answers, they may answer differently) and to whether it is practicable (e.g. if the hospital or general practice can actually handle the dispatch and return of questionnaires). (Recommendation based on evidence from primary studies, findings from previous reviews and expert opinion.)

Saliency

- As far as possible, ensure the saliency (relevance and interest) of the survey topic to the study population. Fortunately, surveys on health-related topics are generally perceived to be highly salient. (Recommendation based on evidence from primary studies and findings from previous reviews.)

Incentives

- If ethical and budgetary constraints allow, consider the use of enclosed financial incentives. In making the choice, the most relevant cost to consider is the projected cost per returned questionnaire; will the likely additional yield in responses outweigh the additional cost of providing the incentive? Note also that incentives are often regarded as unethical in health research, and grant-awarding bodies tend to disapprove of the practice.[3] (Recommendation

based on evidence from primary studies, findings from previous reviews and expert opinion.)

Recommendations derived solely from theories of respondent behaviour, previous literature reviews and/or expert opinion

Timing of survey

- If possible, avoid the month of December in conducting postal surveys. Depending on the survey topic and the study population, avoidance of the peak holiday months (July and August) may also be advisable.

Anonymity/confidentiality

- Provide appropriate assurances of confidentiality, on the questionnaire itself and in the covering letter. Clarify what confidentiality means in the context of the specific survey (generally that only the research team will be able to link the numbered questionnaire to a named individual and that individual responses will not be revealed to a third party without the explicit permission of the respondent concerned).
- If totally unidentifiable questionnaires are deemed to be necessary, consider the use of a numbered (and therefore identifiable) postcard to be returned under separate cover. This will facilitate the use of reminders, although not record linkage.

Covering letter: style and content

- Keep the covering letter short and use language that is appropriate to the target recipients. If extensive or detailed information needs to be given, consider including a separate information sheet.
- Include contact details for the research organisation and ensure that all those who are likely to receive enquiries are adequately briefed.

Postage rates and types

- Always include a prepaid and addressed return envelope. (Recommendation based on expert opinion; no relevant studies identified.)
- Add a return address to the outside of the outgoing envelope to facilitate the return of undeliverable mail. (Recommendation based on the accumulated experience of the review team and advice from the Royal Mail; no relevant studies identified.)

Provision of feedback and results

- Providing feedback to study respondents is probably unnecessary in surveys of the general public, but it may be appropriate in surveys of health professionals. (Recommendation based on the accumulated experience of the review team; evidence from primary studies shows little effect, but does not relate to surveys of special populations.)

Chapter 9

Summary of recommendations for future research

Mode of administration (chapter 3)

With the growing availability of and interest in information technology, priority should be given to comparative studies of traditional versus computer-assisted approaches, of different computer-assisted approaches with each other, and of mixed-mode approaches, for example:

- CATI and CAPI versus CASI
- traditional modes of data entry (data keying) from paper-based questionnaires versus electronic scanning of questionnaires (OMR and OCR)
- traditional keyboard entry for computer-assisted questionnaires versus more novel techniques such as touch-screen and light-pen data entry
- web-based delivery of questionnaires (particular issues here would be how to define and determine the underlying population and how to control for the same individual submitting multiple questionnaires)
- incorporation of traditional or CASI segments into interviewer-administered surveys (e.g. to gather data on sensitive topics).

Particular attention should be paid to the relative merits of different modes of administration in surveys:

- of special populations (e.g. older people, ethnic communities, hearing-impaired people, motor-impaired people, health professionals)
- on sensitive topics (e.g. sexual behaviour, drug and alcohol use).

Future comparative studies of different modes of administration should use multiple outcome measures, including:

- the quantity of response (non-contact, ineligibility, refusal and instrument response rates; item non-response rates)
- the quality of response (non-response bias; validity, reliability and distribution of responses)
- resource implications (time to respond; cost per completed questionnaire).

Question wording and sequencing (chapter 4)

Although some aspects of expert opinion with regard to question wording, question ordering and the construction of response categories have not been subjected to experimental manipulation, their sense is self-evident and further investigation is unlikely to be fruitful. For example, there is no reason to believe that experts' recommendations about avoiding ambiguity in question wording would be refuted through comparisons of ambiguous and unambiguous questions.

Some other aspects of question wording, question ordering and the construction of response categories are, however, ripe for further investigation; priorities are set out below. For the most part, the authors recommend prioritising those aspects of question and response construction that have not been extensively studied to date. However, they also suggest that it will be important to test whether effects that have already been demonstrated in one context and with one mode of survey administration are also found in other settings and with other modes of administration, and to replicate new investigations across different modes of administration. In particular, comparisons between interviewer-administered and self-completion approaches are warranted because theories of response formulation, previous research and expert opinion suggest that different types of response bias may occur under the two modes.[52,84,117,121]

Study designs in which respondents are allocated randomly to different versions of a questionnaire (e.g. 5- versus 7-point response scales) will be appropriate in examining the effects of question wording, ordering and response category construction. However, split-half designs, in which each questionnaire contains a mix (again, randomly assigned) of items could also be considered.

As well as quantitative experimental research, qualitative methods, in particular cognitive testing techniques,[134–137] will be appropriate in assessing how respondents comprehend questions and formulate their responses (appendix 1).

Key measures for research into question construction

In comparisons of aspects of question construction, one key measure will be the validity of responses, in other words, whether the question is truly measuring what it purports to measure. Another key indicator will be the reliability of responses, that is, whether the question or questionnaire is measuring things in a consistent or reproducible way. The assessment of validity and reliability is discussed in greater detail in appendix 1. These topics are also discussed in a number of key texts and articles (e.g.[22,138,139]). In addition to validity and reliability, the precision and discriminatory power of questions and their associated response categories need to be considered. Questions to which the vast majority of respondents choose the same response category are unlikely to be discriminating.[22] An examination of the distribution of responses across the response categories, using measures of spread and skewness, is advisable.

Priorities for research

Further research is required on all three main areas covered by this review. Within each area, the recommendations for research are presented in priority order below.

Question wording

- Questions on the frequency and periodicity of behaviour are the key to many health-related surveys. Further research into time-framing of questions (e.g. "1 month" versus "3 months") and of different quantifiers for time-related questions (e.g. "how many times" versus "how often") is therefore indicated. There may be trade-offs between validity, reliability and discriminatory power of the different quantifiers, and it will be important to take account of this in analysing data from such investigations.
- Studies of aided-recall techniques (e.g. bounded recall) for memory questions are recommended; no research on this topic was identified.
- Comparative studies of the different methods suggested by Sudman and Bradburn[7] (described in the first part of chapter 4) for deliberately loading threatening or sensitive questions, in order to obtain more valid responses, should be carried out. No research was identified on this topic.
- Conventional wisdom suggests that a mix of positively and negatively worded statements should be used in measuring attitudes, but the limited evidence from the one identified study on this topic[89] concluded that the inclusion of negative items in attitudinal questionnaires may impair rather than increase the validity of survey results. Further research into the impact of mixing positive and negative statements is therefore recommended.
- Limited evidence was found concerning the impact of filter questions. The authors therefore advocate comparisons of the inclusion and exclusion of filter questions and suggest that these should focus on: filtering out respondents with no preformulated opinions before asking detailed questions about attitudes; using filter questions to avoid asking detailed questions of people who have no knowledge of a topic; and filtering out those respondents who have never engaged in the investigated form of behaviour.

Question sequencing

- Theories of respondent behaviour suggest that question ordering effects may be reduced in self-completion questionnaires (because the respondent has the opportunity to preview all the questions before responding), but empirical evidence on this topic is limited. The authors therefore advocate that research into the effect of question ordering should concentrate on self-completion questionnaires. Theories of respondent behaviour suggest that ordering effects are most marked in respect of attitudinal questions,[115,116] so it is suggested here that these should be the first priority in future investigations.
- Social desirability bias may occur when behaviour questions are asked after knowledge questions on a related topic (e.g. questions on personal dietary behaviour after items on knowledge of good eating practice), so comparisons of the relative position of these sets of questions are warranted. No existing studies on this specific aspect of question ordering were located.
- The apparent relevance and "ease of answering" of opening questions may influence the decision to respond,[1,140] so comparisons of more and less salient opening items are indicated.

Response categories

- The ordering of response categories may lead to response bias (both recency and primacy effects[52,84]), therefore further comparative studies of alternative ordering are desirable; this is particularly true for questions on sensitive topics, where it has been suggested that the categories should be ordered from the least to the most socially desirable.[7]
- Recency effects appear to be more common in interviewer-administered surveys.[52,84] The authors therefore recommend studies on ways of minimising such effects (e.g. the use of prompt cards; whether multiple-step approaches are any more effective than single-step methods; what techniques can be used in telephone surveys).

- Sudman and Bradburn[7] have suggested that analogue scales (e.g. ladders, clocks, thermometers) may be effective for numerical scales with many points. No studies of such approaches were identified and the authors recommend that they should be compared with more conventional numerical scales.
- It has been suggested that increased precision may be achieved through the use of seven rather than five response categories,[36] especially in Likert-type scales, and there is some evidence for this.[141] There is little evidence, however, for increased enhancement of precision beyond seven categories. Further research into the reliability and discriminatory power of five- versus seven-point (or more finely graded) scales is recommended.
- Findings from identified comparative studies of the labelling of all scale points compared with attaching verbal descriptors to end-points only are equivocal.[125,130] Further research into this topic is therefore desirable.

Questionnaire appearance (chapter 5)

Issues of questionnaire format and appearance have been under-researched to date. The time is therefore ripe for studies on the impact of questionnaire design. The authors of this review recommend that, in designing studies comparing aspects of questionnaire design, researchers should draw on theories of perception, pattern recognition and cognition[145] and should seek to test out the common recommendations of survey experts.[1,7,13,40]

Comparative studies should use multiple outcome measures, including:

- the quantity of response (instrument response rates; item non-response rates)
- the quality of response (non-response bias; validity, reliability and distribution of responses)
- resource implications (cost per completed questionnaire).

In addition to the quantifiable measures identified above, cognitive testing[134–137] of how respondents react to different design features should be employed.

Priorities for research

No evidence was identified on the following aspects of questionnaire appearance: double- versus single-sided printing; placement of questions within pages; print details; cover design; methods of identifying questions; vertical versus horizontal response formats; placement of headings and codes for response categories; identification of skip patterns; and nature, placing or format of instructions. However, the authors regard some of these topics (e.g. double- versus single-sided printing) as of low priority for future research. Below are presented the authors' priorities for research, in order of importance:

- research into the possibility that design principles for computer-assisted surveys (OMR and OCR technology, web-based questionnaires, which, as noted in the recommendations for practice, are likely to assume growing importance in the future) may differ from the principles espoused for paper-based questionnaires
- further testing of the impact of questionnaire length, particularly when the topic may be perceived as less salient
- formal testing of vertical versus horizontal response formats for multiple-choice questions
- studies of the relative placement of headings, response category descriptors and codes
- studies of verbal and graphical methods (including the use of colour contrast and different typefaces) to aid "navigation" through the questionnaire
- studies of the placement and format of instructions for interviewers, respondents and data processors.

Enhancing response rates (chapter 6)

Methods of enhancing response rates have already been extensively researched.[27] However, much of this study has been in the fields of social, educational and market research. A high priority for research, therefore, should be to examine whether techniques previously shown to enhance response rates in non-health-related surveys are also effective in stimulating responses to health surveys. Given that response rates to surveys on health-related topics are generally higher, it is possible that there may be a ceiling effect. The authors consider that research should also focus on whether effective methods of enhancing response rates are common to health surveys of general populations, special patient or consumer groups, and health professionals.

In designing primary studies, researchers should seek to challenge expert opinion, as summarised in the frameworks provided by Dillman[1] and by

Brown and colleagues,[140] and to test theories of respondent behaviour. Experimental manipulation of aspects of survey design and administration will be best carried out in a "real world" setting, by "piggy-backing" an experiment on to a real survey, rather than by creating an artificial situation and carrying out a survey simply for the sake of testing one or more factors hypothesised to affect response rates. In manipulating factors, care should be taken to use a realistic combination (e.g. to avoid combining a high degree of personalisation with an assurance of complete anonymity). In analysis, the interaction between manipulated factors, as well as the main effects of each individual factor, should be examined.

Comparative studies should use multiple outcome measures; there is little point in boosting the quantity of response (i.e. response rates) if this is at the expense of the quality of response (e.g. increased non-response bias, less complete responses, greater response bias). Moreover, more intensive approaches, although effective, may not be cost-effective; a key outcome variable should be the cost per usable questionnaire.

The priority order for further research will depend on the study population. For example, the authors believe that research into modes of contact and follow-up is particularly relevant in respect of surveys of health professionals, while studies with partial anonymity are more important for patient populations, especially in surveys on sensitive topics.

Priorities for research

- Mode of contact, especially for reminders: Anecdotally, it has been suggested that telephone reminders (perhaps with an offer to complete the questionnaire as a telephone interview) may be particularly appropriate in surveys of professional groups.
- Follow-up messages: Further research is particularly indicated into whether stressing either or both of (1) the importance of the individual's response, and (2) response rates to date, is beneficial in stimulating response rates.
- Partial anonymity: Sudman suggested[181] a comparison of identifiable (numbered) questionnaires with unidentifiable questionnaires accompanied by an identifiable postcard to be posted back separately to indicate that the questionnaire has also been returned.
- Personal delivery/collection of self-completion questionnaires: Limited evidence from one cross-sectional study[263] suggests that this may be a useful method of boosting response rates. Research into whether the potential increase in response quantity and quality outweighs the likely additional costs is recommended.
- Personalisation: Although findings from existing studies on the effects of personalisation suggest little benefit, the authors recommend testing whether a personalised letter is more effective than a form letter in situations where the target respondents are "personally" known to the researchers (as would be the case, for example, in a survey of patients by their own GP).
- Incentives: Although personal financial incentives may be regarded as unethical or inappropriate, a promised donation to charity may be more acceptable. Although evidence to date suggests that promises of untargeted charitable donations are not very effective in stimulating response in general surveys, research into whether a promised donation to a relevant charity may be effective in surveys of specific patient or consumer groups is recommended.
- Nature of appeal in covering letters: Further comparisons of "egoistic/self-interest" versus "altruistic/social utility" appeals are recommended, especially in surveys of special patient or consumer groups.
- "Foot-in-door" techniques: Evidence on the effectiveness of these techniques is mixed and further investigation of their value (given that they are resource intensive) is warranted, especially in relation to health surveys.
- Provision of information about the survey/research topic: It is recommended that studies should be conducted on whether providing more detailed information about a research project and how that information is provided (covering letter versus separate information sheet) has an effect on response rates. The cost per returned questionnaire would be an important outcome variable because the inclusion of extra information may have significant resource implications. No existing studies on this topic were found.
- Provision of time cues: There is limited evidence on the effectiveness of this approach[48] and the authors therefore recommend further investigation of specification in the covering letter and/or in the questionnaire itself of the likely time required for completion.
- "Threat" of follow-up and specification of deadlines for return of the questionnaire: Comparative studies on the inclusion of a statement in the covering letter accompanying the original questionnaire indicating that reminders will be sent if the questionnaire is not returned within perhaps 2 weeks are desirable. The effect of specifying a deadline for response

should also be investigated. Speed of response, as well as response rates, should be monitored.
- Timing of survey: In particular, studies are required to verify whether expert advice to avoid July, August and December is borne out in practice. A key outcome variable, in addition to response rates, should be speed of response.

In addition to these primary research studies, the authors also suggest that:

- In all surveys, researchers should attempt to quantify and report the extent and nature of non-response bias, and to analyse whether there are important differences between early and late respondents.
- Reviews of the methods and results of well-designed health-related surveys (e.g.[265,266]) should be carried out as a low-cost and low-key approach to identifying good practice in the conduct of health surveys, although caution should be exercised in generalising from surveys of specific populations.
- Qualitative research, including cognitive interviewing,[134–137] should be carried out with both lay and professional groups to investigate barriers to and facilitators of participation in surveys, including motivational factors.

Chapter 10

Trajectory of the knowledge base

As already noted, research into aspects of survey design and administration to date has been quite haphazard. It appears that individual researchers identify topics of personal or organisational interest and either design experimental interventions specifically to address these concerns or incorporate such experiments (or *post hoc* analyses) into ongoing surveys. Through this review of the literature, the authors acquired no sense of a focused approach to this research in the form of international, national or even local (at the level of the individual survey organisation) agendas. Despite many researchers making recommendations that their experiments should be repeated in other settings, or under different modes of survey administration, there was scant evidence of replication of studies to check the generalisability of findings.

For these reasons, it is the authors' opinion that, for most of the areas covered by this review, new knowledge (from primary research studies) will accumulate relatively slowly. Of course, further studies on the effectiveness of, for example, incentives, will continue to be carried out, but it is debatable how much the findings from these studies will add to the sum of current knowledge.

Through this review the authors have become aware, however, of ongoing programmes of research into certain aspects of survey methodology, particularly concerning issues of the wording and ordering of questions and associated response categories. Prominent in such research are national survey research organisations such as the National Opinion Research Centre and the Survey Methods Centre in the USA, and the National Centre for Social Research in the UK. The findings from these research programmes tend to be published initially in the "grey literature" (reports and technical bulletins published by the organisations concerned), followed by dissemination in peer-reviewed journals (e.g. *Public Opinion Quarterly* and *Survey Methods Bulletin*). The authors suggest that the research programmes of these specialist research organisations should be monitored.

Another growing area of interest, and one in which peer-reviewed articles reporting on high-quality studies are beginning to appear, is that of computer-assisted survey administration, including the use of the Internet[63] as a method of questionnaire delivery. The authors believe that, with growing interest in these emerging technologies, there will be a rapid accumulation of evidence in this area, and recommend this as a priority for future reviews.

Finally, there is now increasing interest in applying and testing principles of cognition[134–137] in gaining a deeper understanding of the way in which survey respondents react to and respond to questions, questionnaires and related documents. From this ongoing research it is possible that theories of respondent behaviour will be honed and refined.

In the light of their experience in undertaking this current review, the authors believe that repeating/updating the entire review in perhaps 5 years would not be a very efficient use of resources. Instead, a more targeted approach is recommended, involving perhaps a rolling programme of updates. Initially, the focus should be on those areas of survey design and administration for which little evidence to date was identified (for example, e-mail surveys, aspects of questionnaire layout). In looking for evidence from primary studies on these "new" topics, the search strategy should include both the "grey" and the published literature from the fields of social, educational and market research (because, as already identified, much of the novel research tends to be initiated in these fields) as well as from health-related research, and without geographical restriction. For those areas of survey design and administration on which there is already a considerable body of accumulated knowledge from primary research studies in a range of settings (e.g. the use of incentives in stimulating response rates), the emphasis should be on seeking confirming or disconfirming evidence from health-related studies.

Acknowledgements

This review was commissioned by the NHS R&D Health Technology Programme.

The authors acknowledge the invaluable assistance of the following colleagues, without whom this work would have been impossible: Mrs Linda Duckworth, for literature retrieval, database input and secretarial support; Ms Rowan Standish, for secretarial support; Mrs Barbara Ingman for technical support; Ms Carol Riccalton, for advice on literature searching; and Mrs Sylvia Hudson, for researching postal options (appendix 1).

The opinions expressed are those of the authors alone.

References

1. Dillman DA. Mail and telephone surveys: the total design method. New York: Wiley; 1978.
2. Fishbein M. A consideration of beliefs and their role in attitude measurement. In: Fishbein M, editor. Readings in attitude theory and measurement. New York: Wiley; 1967. p. 257–66.
3. Bowling A. Research methods in health: investigating health and health services. Buckingham: Open University Press; 1997.
4. Sackett DL. Bias in analytic research. *Journal of Chronic Diseases* 1979;**32**:51–63.
5. Moser CA, Kalton G. Survey methods in social investigation. 2nd ed. Aldershot: Gower; 1971.
6. Bennett AE, Ritchie R. Questionnaires in medicine: a guide to their design and use. London: Oxford University Press; 1975.
7. Sudman S, Bradburn N. Asking questions: a practical guide to questionnaire design. San Francisco, CA: Jossey-Bass; 1982.
8. Bradburn N, Sudman S. Improving interview method and questionnaire design. San Francisco: Jossey-Bass; 1979.
9. Abramson JH. Survey methods in community medicine. Edinburgh: Churchill Livingstone; 1990.
10. Advisory Group on Health Technology Assessment. Assessing the effects of health technologies. London: Department of Health; 1993.
11. Franklin B, Osborne H. Research methods: issues and insights. Belmont, CA: Wadsworth; 1971.
12. Nay-Brock RM. A comparison of the questionnaire and interviewing techniques in the collection of sociological data. *Australian Journal of Advanced Nursing* 1984;**2**:14–23.
13. Oppenheim AN. Questionnaire design, interviewing and attitude measurement. 2nd ed. London: Pinter; 1992.
14. Fowler FJ. Survey research methods. 2nd ed. Beverly Hills, CA: Sage; 1993.
15. Salant P, Dillman DA. How to conduct your own survey. New York: Wiley; 1994.
16. Mangione TW. Mail surveys – improving the quality. Thousand Oaks, CA: Sage; 1995.
17. Fink A. The survey kit. Thousand Oaks, CA: Sage; 1995.
18. Crombie IK, Davies HTO. Research in health care: design, conduct and interpretation of health services research. Chichester: Wiley; 1996.
19. Øvretveit J. Evaluating health interventions. Buckingham: Open University Press; 1998.
20. Fink A, Kosecoff J. How to conduct surveys: a step-by-step guide. Thousand Oaks, CA: Sage; 1998.
21. Dillman DA. Mail and internet surveys: the tailored design method. 2nd ed. New York: Wiley; 2000.
22. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. Oxford: Oxford University Press; 1989.
23. Murphy EA. The logic of medicine. Baltimore, MD: Johns Hopkins University Press; 1976.
24. Goyder J. Survey errors and survey costs. New York: Wiley; 1989.
25. Lynn P. Quality and error in self-completion surveys. *Survey Methods Centre Newsletter* 1996;**16**:4–9.
26. Sudman S, Bradburn N. Response effects in surveys: a review and synthesis. Chicago, IL: Aldine; 1974.
27. Dickinson JR. The bibliography of marketing research methods. Lexington, KT: Heath and Company; 1986.
28. Hutton JL, Ashcroft R. What does "systematic" mean for reviews of methods? In: Black N, Brazier J, Fitzpatrick R, Reeves B, editors. Methods for health care services research. London: British Medical Journal Books; 1998. p. 249–54.
29. Edwards SJL, Lilford RJ, Kiauka S. Different types of systematic review in health services research. In: Black N, Brazier J, Fitzpatrick R, Reeves B, editors. Methods for health care services research. London: British Medical Journal Books; 1998. p. 255–9.
30. Oxman AD. The Cochrane Collaboration handbook: preparing and maintaining systematic reviews. Oxford: Cochrane Collaboration; 1996.
31. Murphy E, Dingwall R, Greatbach D, Parker S, Watson P. Qualitative research methods in health technology assessment: a review of the literature. *Health Technology Assessment* 1998;**2**(16).
32. McDowell I, Newall C. Measuring health: a guide to rating scales and questionnaires. Oxford: Oxford University Press; 1987.
33. Wilkin D, Hallam L, Doggett M-A. Measures of need and outcome for primary health care. Oxford: Oxford University Press; 1992.
34. Bowling A. Measuring disease. Buckingham: Open University Press; 1995.

35. Bowling A. Measuring health: a review of quality of life measurement scales. 2nd ed. Buckingham: Open University Press; 1997.

36. Fitzpatrick R, Davey C, Buxton MJ, Jones DR. Evaluating patient-based outcome measures for use in clinical trials. *Health Technology Assessment* 1998;**2**(14).

37. Murphy MK, Black NA, Lamping DL, McKee CM, Sanderson CFB, Askham J, *et al.* Consensus development methods, and their use in clinical guideline development. *Health Technology Assessment* 1998;**2**(3).

38. Edwards SJL, Lilford RJ, Braunholtz DA, Jackson JC, Hewison J, Thornton J. Ethical issues in the design and conduct of randomised controlled trials. *Health Technology Assessment* 1998;**2**(15).

39. Cook TD, Campbell DT. Quasi-experimentation. Boston, MA: Houghton Mifflin; 1979.

40. Bourque LB, Fielder EP. How to conduct self-administered and mail surveys. Thousand Oaks, CA: Sage; 1995.

41. Sheldon TA, Song F, Davey Smith G. Critical appraisal of the medical literature: how to assess whether health-care interventions do more good than harm. In: Drummond MF, Maynard A, Wells N, editors. Purchasing and providing cost-effective health care. Edinburgh: Churchill Livingstone; 1993. p. 31–48.

42. Stock WA. Systematic coding for research synthesis. In: Cooper H, Hedges LV, editors. The handbook of research synthesis. New York: Russell Sage Foundation; 1994. p. 125–38.

43. Wortman PM. Judging research quality. In: Cooper H, Hedges LV, editors. The handbook of research synthesis. New York: Russell Sage Foundation; 1994. p. 98–109.

44. Gardner MJ, Altman DG. Statistics with confidence – confidence intervals and statistical guidelines. London: British Medical Journal Books; 1989.

45. Thornton A, Lee P. Publication bias in meta-analysis: its causes and consequences. *Journal of Clinical Epidemiology* 2000;**53**:207–16.

46. Song F, Eastwood AJ, Gilbody S, Duley L, Swadi H. Publication and related biases. *Health Technology Assessment* 2000;**4**(10).

47. Cartwright A. Some experiments with factors that might affect the response of mothers to a postal questionnaire. *Statistics in Medicine* 1986;**5**:607–17.

48. Hornik J. Time cue and time perception effect on response to mail surveys. *Journal of Marketing Research* 1981;**18**:243–8.

49. McAvoy BR, Kaner EFS. General practice postal surveys: a questionnaire too far? *British Medical Journal* 1996;**313**:732–3.

50. de Vaus DA. Surveys in social research. 3rd ed. London: UCL Press; 1991.

51. Morton Williams J. Interviewer approaches. Aldershot: Dartmouth; 1993.

52. Krosnick JA. The threat of satisficing in surveys: the shortcuts respondents take in answering questions. *Survey Methods Centre Newsletter* 2000;**20**:4–8.

53. Hyman HH, Cobb WJ, Feldman JJ, Hart CW, Stember CH. Interviewing in social research. Chicago, IL: University of Chicago Press; 1954.

54. Williams JA. Interviewer–respondent interaction: a study in bias in the information interview. *Sociometry* 1964;**27**:338–52.

55. Schuman H, Converse JM. The effect of black and white interviewers on black responses in 1968. *Public Opinion Quarterly* 1971;**35**:44–68.

56. Welch S, Comer J, Steinman M. Interviewing in a Mexican–American community: an investigation of some potential sources of response bias. *Public Opinion Quarterly* 1973;**37**:115–26.

57. Hatchett S, Schuman H. White respondents and race-of-interviewer effects. *Public Opinion Quarterly* 1975;**39**:523–8.

58. Schaeffer NC. Evaluating race-of-interviewer effects in a national survey. *Sociological Methods and Research* 1980;**8**:400–19.

59. Weeks MF, Moore RP. Ethnicity-of-interviewer effects on ethnic respondents. *Public Opinion Quarterly* 1981;**45**:245–9.

60. Frey JH, Oishi SM. How to conduct interviews by telephone and in person. Thousand Oaks, CA: Sage; 1995.

61. Mishra SJ, Dooley D, Catalano R, Serxner S. Telephone health surveys: potential bias from noncompletion. *American Journal of Public Health* 1993;**83**:94–9.

62. Harlow BL, Rosenthal JF, Ziegler RG. A comparison of computer-assisted and hard copy telephone interviewing. *American Journal of Epidemiology* 1985;**122**:335–40.

63. Mehta R, Siviadis E. Comparing response rates and response content in mail versus electronic mail surveys. *Journal of the Market Research Society* 1995;**37**:429–39.

64. Hinkle AL, King GD. A comparison of three survey methods to obtain data for community mental health program planning. *American Journal of Community Psychology* 1978;**6**:389–97.

65. Talley JE, Barrow JC, Fulkerson KF, Moore CA. Conducting a needs assessment of university psychological services: a campaign of telephone and mail strategies. *Journal of American College Health* 1983;**32**:101–3.

66. McHorney CA, Kosinski M, Ware JE. Comparisons of the costs and quality of norms for the SF-36 health survey collected by mail versus telephone interview: results from a national survey. *Medical Care* 1994;**32**:551–67.

67. Pederson LL, Baskerville JC, Ashley MJ, Lefcoe NM. Comparison of mail questionnaire and telephone interview as data gathering strategies in a survey of attitudes toward restrictions on cigarette smoking. *Canadian Journal of Public Health* 1994;**76**:179–82.

68. Newton RR, Prensky D, Schuessler K. Form effect in the measurement of feeling states. *Social Science Research* 1982;**11**:301–17.

69. Nederhof AJ. Visibility of response as a mediating factor in equity research. *Journal of Social Psychology* 1984;**122**:211–15.

70. Oei TI, Zwart FM. The assessment of life events: self-administered questionnaire versus interview. *Journal of Affective Disorders* 1986;**10**:185–90.

71. Cartwright A. Interviews or postal questionnaires? Comparisons of data about women's experiences with maternity services. *Milbank Quarterly* 1988;**66**:172–89.

72. Liefeld JP. Response effects in computer-administered questioning. *Journal of Marketing Research* 1988;**25**:405–9.

73. Boekeloo BO, Schiavo L, Rabin DL, Conlon RT, Jordan CS, Mundt DJ. Self-reports of HIV risk factors by patients at a sexually transmitted disease clinic: audio vs. written questionnaires. *American Journal of Public Health* 1994;**84**:754–60.

74. Jordon LA, Marcus AC, Reeder LG. Response styles in telephone and household interviewing: a field experiment. *Public Opinion Quarterly* 1980;**44**:210–22.

75. Quinn RP, Gutek BA, Walsh JT. Telephone interviewing: a reappraisal and a field experiment. *Basic and Applied Social Psychology* 1980;**1**:127–53.

76. Aneshensel CS, Frerichs RR, Clark VA, Yokopenic PA. Telephone versus in person surveys of community health status. *American Journal of Public Health* 1982;**72**:1017–21.

77. Fenig S, Levav I, Kohn R, Yelin N. Telephone vs. face-to-face interviewing in a community psychiatric survey. *American Journal of Public Health* 1993;**83**:896–8.

78. Allen DF. Computers versus scanners: an experiment in nontraditional forms of survey administration. *Journal of College Student Personnel* 1987;**28**:266–73.

79. Higgins CA, Dimnik TP, Greenwood HP. The DISKQ survey method. *Journal of the Market Research Society* 1987;**29**:437–45.

80. Helgeson JG, Ursic ML. The decision process equivalency of electronic versus pencil-and-paper data collection methods. *Social Science Computer Review* 1989;**7**:296–310.

81. Groves RM, Mathiowetz NA. Computer assisted telephone interviewing: effects on interviewers and respondents. *Public Opinion Quarterly* 1984;**48**:356–69.

82. Miller PV. Alternative question forms for attitude scale questions in telephone interviews. *Public Opinion Quarterly* 1984;**48**:766–78.

83. Belson WA. The design and understanding of survey questions. Aldershot: Gower; 1981.

84. Schuman H, Presser S. Questions and answers in attitude surveys: experiments on question form, wording and content. New York: Academic Press; 1981.

85. Converse JM, Presser S. Survey questions: handcrafting the standardised questionnaire. Beverly Hills, CA: Sage; 1986.

86. Fowler FJ. Improving survey questions: design and evaluation. Thousand Oaks, CA: Sage; 1995.

87. Markum RA. Assessment of the reliability of and the effect of neutral instructions on the symptom ratings on the Moos Menstrual Distress Questionnaire. *Psychosomatic Medicine* 1976;**38**:163–72.

88. Salvendy G. Some variables which affect mail survey response rate. *Studia Psychologica* 1976;**18**:250–2.

89. Schriesheim CA, Hill KD. Controlling acquiescence response bias by item reversals: the effect on questionnaire validity. *Educational and Psychological Measurement* 1981;**41**:1101–14.

90. Bishop GF, Oldendick RW, Tuchfarber AJ. Effects of presenting one versus two sides of an issue in survey questions. *Public Opinion Quarterly* 1982;**46**:69–85.

91. Bishop GF, Oldendick RW, Tuchfarber AJ. Effects of filter questions in public opinion surveys. *Public Opinion Quarterly* 1983;**47**:528–46.

92. Bishop GF, Tuchfarber AJ, Oldendick RW. Opinions on fictitious issues: the pressure to answer survey questions. *Public Opinion Quarterly* 1986;**50**:240–50.

93. Schuman H, Ludwig J, Krosnick JA. The perceived threat of nuclear war, salience, and open questions. *Public Opinion Quarterly* 1986;**50**:519–36.

94. Blair E, Burton S. Cognitive processes used by survey respondents to answer behavioral frequency questions. *Journal of Consumer Research* 1987;**14**:280–8.

95. Larsen JD, Mascharka C, Toronski C. Does the wording of the question change the number of headaches people report on a health questionnaire? *Psychological Record* 1987;**37**:423–7.

96. Barnes JH, Dotson MJ. The effect of mixed grammar chains on response to survey questions. *Journal of Marketing Research* 1989;**26**:468–72.

97. Smith ER, Squire P. The effects of prestige names in question wording. *Public Opinion Quarterly* 1990;**54**:97–116.

98. Brown CH. Wittgensteinian linguistics. Amsterdam: Mouton; 1974.

99. Chomsky N. Aspects of the theory of syntax. Cambridge: MIT Press; 1965.

100. Cannell CF, Oksenberg L, Converse JM. Striving for response accuracy. *Journal of Marketing Research* 1977;**14**:306–15.

101. Sudman S, Ferber R. A comparison of alternative procedures for collecting consumer expenditure data for frequently purchased products. *Journal of Marketing Research* 1974;**11**:128–35.

102. Wind Y, Lerner D. On the measurement of purchase data: surveys versus purchase diaries. *Journal of Marketing Research* 1979;**16**:39–47.

103. Neter J, Waksberg J. A study of response errors in expenditure data from household interviews. *Journal of the American Statistical Association* 1964;**59**:18–55.

104. Bradburn NM, Mason WM. The effect of question order on responses. *Journal of Marketing Research* 1964;**1**:57–61.

105. Smith TW. Condition order effects (GSS technical report no. 33). Chicago, IL: National Opinion Research Centre; 1982.

106. Hippler H-J, Schwarz N. Response effects in surveys. In: Hippler H-J, Schwarz N, Sudman S, editors. Social information processing and survey methodology. New York: Springer-Verlag; 1987. p. 102–22.

107. Schuman H, Presser S, Ludwig J. Context effects on survey responses to questions about abortion. *Public Opinion Quarterly* 1981;**45**:216–23.

108. Colasanto D, Singer E, Rogers TF. Context effects on responses to questions about AIDS. *Public Opinion Quarterly* 1992;**56**:515–18.

109. Barry MJ, Walker-Corkery E, Chang Y, Tyll LT, Cherkin DC, Fowler FJ. Measurement of overall and disease-specific health status: does the order of questionnaires make a difference? *Journal of Health Services Research and Policy* 1996;**1**:20–7.

110. Jones WH, Lang JR. Sample composition bias and response bias in a mail survey: a comparison of inducement methods. *Journal of Marketing Research* 1980;**17**:69–76.

111. McFarland SG. Effects of question order on survey responses. *Public Opinion Quarterly* 1981;**45**:208–15.

112. Sigelman L. Question-order effects on presidential popularity. *Public Opinion Quarterly* 1981;**45**:199–207.

113. Schuman H, Kalton G, Ludwig J. Context and contiguity in survey questionnaires. *Public Opinion Quarterly* 1983;**47**:112–15.

114. Spector PE, Michaels CE. A note on item order as an artifact in organizational surveys. *Journal of Occupational Psychology* 1983;**56**:35–6.

115. Tourangeau R, Rasinski KA, Bradburn N, D'Andrade R. Belief accessibility and context effects in attitude measurement. *Journal of Experimental Social Psychology* 1989;**25**:401–21.

116. Tourangeau R, Rasinski KA, Bradburn N, D'Andrade R. Carryover effects in attitude surveys. *Public Opinion Quarterly* 1989;**53**:495–524.

117. Ayidiya SA, McClendon MJ. Response effects in mail surveys. *Public Opinion Quarterly* 1990;**54**:229–47.

118. Roberson MT, Sundstrom E. Questionnaire design, return rates, and response favorableness in an employee attitude questionnaire. *Journal of Applied Psychology* 1990;**75**:354–7.

119. Tenvergert E, Gillespie MW, Kingma J, Klasen H. Abortion attitudes, 1984–1987–1988: effects of item order and dimensionality. *Perceptual and Motor Skills* 1992;**74**:627–42.

120. Serdula MK, Mokdad AH, Pamuk ER, Williamson DF, Byers T. Effects of question order on estimates of the prevalence of attempted weight loss. *American Journal of Epidemiology* 1995;**142**:64–7.

121. Bishop GF, Hippler H-J, Schwarz N, Strack F. A comparison of response effects in self-administered and telephone surveys. In: Groves RM, editor. Telephone survey methodology. New York: Wiley; 1988. p. 321–40.

122. Fletcher A, Gore S, Jones D, Fitzpatrick R, Spiegelhalter D, Cox D. Quality of life measures in health care: II – design, analysis and interpretation. *British Medical Journal* 1992;**305**:1145–8.

123. Poe GS, Seeman I, McLaughlin J, Mehl E. "Don't know" boxes in factual questions in a mail questionnaire: effects on level and quality of response. *Public Opinion Quarterly* 1988;**52**:212–22.

124. Stem DE, Lamb CW, MacLachlan DL. Remote versus adjacent scale questionnaire designs. *Journal of the Market Research Society* 1978;**20**:3–13.

125. Frisbie DA, Brandenburg DC. Equivalence of questionnaire items with varying response formats. *Journal of Educational Measurement* 1979;**16**:43–8.

126. Edvardsson B. Effect of reversal of response scales in a questionnaire. *Perceptual and Motor Skills* 1980;**50**:1125–6.

127. Hawkins DI, Coney KA. Uninformed response error in survey research. *Journal of Marketing Research* 1981;**18**:370–4.

128. Israel GD, Taylor CL. Can response order bias evaluations? *Evaluation and Program Planning* 1990;**13**:365–71.

129. Swan JE, Epley DE. Completion and response rates for different forms of income questions in a mail survey. *Perceptual and Motor Skills* 1981;**52**:219–22.

130. Lam TC, Klockars AJ. Anchor point effects on the equivalence of questionnaire items. *Journal of Educational Measurement* 1982;**19**:317–22.

131. Trice AD, Dolan MS. Hotel ratings: V. effects of format and survey length. *Psychological Reports* 1985;**56**:176–8.

132. Bishop GF. Experiments with the middle response alternative in survey questions. *Public Opinion Quarterly* 1987;**51**:220–32.

133. Wandzilak T, Ansorge CJ, Potter G. Utilizing "undecided" option with Likert items: associated measurement problems. *International Journal of Sport Psychology* 1987;**18**:51–8.

134. Campanelli PC, Martin EA, Rothgeb JM. The use of respondent and interviewer debriefing studies as a way to study response errors in survey data. *The Statistician* 1991;**40**:253–64.

135. Oksenberg L, Cannell CF, Kalton G. New strategies for pretesting survey questions. *Journal of Official Statistics* 1991;**7**:349–65.

136. Schwarz N, Bradburn NM, Sudman S. Thinking about answers: the application of cognitive processes to survey methodology. San Francisco, CA: Jossey-Bass; 1996.

137. Schwarz N, Sudman S. Answering questions: methodology for determining cognitive and communicative processes in survey research. San Francisco, CA: Jossey-Bass; 1996.

138. Tulsky DS. An introduction to test theory. *Oncology* 1990;**4**:43–8.

139. Litwin MS. How to measure survey reliability and validity. Thousand Oaks, CA: Sage; 1995.

140. Brown TL, Decker DJ, Connelly NA. Response to mail surveys on resource-based recreation topics: a behavioral model and an empirical analysis. *Leisure Sciences* 1989;**11**:99–110.

141. Avis NE, Smith KW. Conceptual and methodological issues in selecting and developing quality of life measures. In: Fitzpatrick R, editor. Advances in medical sociology. London: JAI Press; 1994. p. 255–80.

142. Thibaut JW, Kelley HH. The social psychology of groups. New York: Wiley; 1959.

143. Homans GC. Social behavior: its elementary forms. New York: Harcourt, Brace and World; 1961.

144. Blau PM. Exchange and power in social life. New York: Wiley; 1964.

145. Jenkins CR, Dillman DA. Towards a theory of self-administered questionnaire design. In: Lyberg L, Biemer P, Collins M, de Leeuw E, Dippo C, Schwarx N, *et al.,* editors. Survey measurement and process quality. New York: Wiley; 1997. p. 165–96.

146. Market Research Society R&DC. Report of the Second Working Party on Respondent Co-operation: 1977–80. *Journal of the Market Research Society* 1981;**23**:3–25.

147. Cannell CF, Kahn RL. Interviewing. In: Lindzey G, Aronson E, editors. The handbook of social psychology – volume II. Reading: Addison-Wesley; 1968. p. 526–95.

148. Bradburn N. Respondent burden. Proceedings of the 2nd Biennial Conference on Health Survey Methods; 1977. Washington: US Department of Health, Education and Welfare and National Centre for Health Services Research.

149. Houston MJ, Ford NM. Broadening the scope of methodological research on mail surveys. *Journal of Marketing Research* 1976;**13**:397–403.

150. Jacoby A. Possible factors affecting response to postal questionnaires: findings from a study of general practitioner services. *Journal of Public Health Medicine* 1990;**12**:131–5.

151. Hansen RA, Robinson LM. Testing the effectiveness of alternative foot-in-the-door manipulations. *Journal of Marketing Research* 1980;**17**:359–64.

152. Layne BH, Thompson DN. Questionnaire page length and return rate. *Journal of Social Psychology* 1981;**113**:291–2.

153. Adams LLM, Gale D. Solving the quandary between questionnaire length and response rate in educational research. *Research in Higher Education* 1982;**17**:231–40.

154. Powers DE, Alderman DL. Feedback as an incentive for responding to a mail questionnaire. *Research in Higher Education* 1982;**17**:207–11.

155. Roszkowski MJ, Bean AG. Believe it or not – longer questionnaires have lower response rates. *Journal of Business and Psychology* 1990;**4**:495–509.

156. Herzog AR, Bachman JG. Effects of questionnaire length on response quality. *Public Opinion Quarterly* 1981;**45**:549–59.

157. Trice AD. Maximizing participation in surveys: hotel ratings VII. *Journal of Social Behavior and Personality* 1986;**1**:137–41.

158. Linsky AS. Stimulating responses to mailed questionnaires: a review. *Public Opinion Quarterly* 1975;**39**:82–101.

159. Kanuk L, Berenson C. Mail surveys and response rates: a literature review. *Journal of Marketing Research* 1975;**12**:440–53.

160. Heberlein TA, Baumgartner R. Factors affecting response rates to mailed questionnaires: a quantitative analysis of the published literature. *American Sociological Review* 1978;**43**:447–62.

161. Goyder JC. Further evidence on factors affecting response rates to mailed questionnaires. *American Sociological Review* 1982;**47**:550–3.

162. Yu J, Cooper H. A quantitative review of research design effects on response rates to questionnaires. *Journal of Marketing Research* 1983;**20**:36–44.

163. Suhre C. Schools over the gangway: an experiment on response improving procedures. *Tijdschrift voor Onderwijsresearch* 1989;**14**:172–80.

164. Jobber D, Sanderson S. The effects of a prior letter and coloured questionnaire paper on mail survey response rates. *Journal of the Market Research Society* 1983;**25**:339–49.

165. Gaskell GD, O'Muircheartaigh CA, Wright DB. Survey questions about vaguely defined events: the effects of response alternatives. *Public Opinion Quarterly* 1994;**58**:241–54.

166. Cochran WG, Chambers SP. The planning of observational studies of human populations. *Journal of the Royal Statistical Society (Series A)* 1965;**128**:234–66.

167. Borg WR, Gall MD. Educational research: an introduction. New York: Longman; 1983.

168. Asch DA, Jedrziewski K, Christiakis NA. Response rates to mail surveys published in medical journals. *Journal of Clinical Epidemiology* 1997;**50**:1129–36.

169. Alwin DF. Making errors in surveys: an overview. *Sociological Methods and Research* 1977;**6**:131–50.

170. Kviz FJ. Towards a standard definition of response rate. *Public Opinion Quarterly* 1977;**41**:265–7.

171. Morton Williams J, Young P. Obtaining the survey interview: an analysis of tape recorded doorstep introductions. *Journal of the Market Research Society* 1987;**29**:35–54.

172. Sosdian CP, Sharp LM. Nonresponse in mail surveys: access failure or respondent resistance. *Public Opinion Quarterly* 1980;**44**:396–402.

173. Norman R. A review of some problems related to the mail questionnaire technique. *Educational and Psychological Measurement* 1948;**8**:234–45.

174. Donald MN. Implications of non-response for the interpretation of mail questionnaire data. *Public Opinion Quarterly* 1960;**24**:99–114.

175. Goyder J. The silent minority: nonrespondents on sample surveys. Cambridge: Polity Press; 1987.

176. Cartwright A. Who responds to postal questionnaires? *Journal of Epidemiology and Community Health* 1986;**40**:267–73.

177. Cartwright A. Professionals as responders: variations in and effects of response rates to questionnaires, 1961–77. *British Medical Journal* 1978;**ii**:1419–21.

178. Hovland EJ, Romberg E, Moreland EF. Nonresponse bias to mail survey questionnaires within a professional population. *Journal of Dental Education* 1980;**44**:270–4.

179. Fiset L, Milgrom P, Tarnai J. Dentists' response to financial incentives in a mail survey of malpractice liability experience. *Journal of Public Health Dentistry* 1994;**54**:68–72.

180. Sudman S. Mail surveys of reluctant professionals. *Evaluation Review* 1985;**9**:349–60.

181. Ward J. General practitioners' experience of research. *Family Practice* 1994;**11**:418–23.

182. Gajraj AM, Faria AJ, Dickinson JR. A comparison of the effect of promised and provided lotteries, monetary and gift incentives on mail survey response rate, speed and cost. *Journal of the Market Research Society* 1990;**32**:141–62.

183. Olivarius NDF, Andreasen AH. Day-of-the-week effect on doctors' response to a postal questionnaire. *Scandinavian Journal of Primary Health Care* 1995;**13**:65–7.

184. Peterson RA, Albaum G, Kerin RA. A note on alternative contact strategies in mail surveys. *Journal of the Market Research Society* 1989;**31**:409–18.

185. Childers TL, Skinner SJ. Theoretical and empirical issues in the identification of survey respondents. *Journal of the Market Research Society* 1985;**27**:39–53.

186. Bagozzi RP. Marketing as exchange. *Journal of Marketing* 1975;**39**:32–9.

187. Kamins MA. The enhancement of response rates to a mail survey through a labelled probe foot-in-the-door approach. *Journal of the Market Research Society* 1989;**31**:273–83.

188. Kelley HH. The process of causal attribution. *American Psychologist* 1973;**28**:107–28.

189. Allen CT, Schewe CD, Wijk G. More on self-perception theory's foot technique in the pre-call/mail survey setting. *Journal of Marketing Research* 1980;**17**:498–502.

190. Nederhof AJ. Effects of preliminary contacts on volunteering in mail surveys. *Perceptual and Motor Skills* 1982;**54**:1333–4.

191. Martin WS, Duncan WJ, Sawyer JC. The interactive effects of four response rate inducements in mail questionnaires. *College Student Journal* 1984;**18**:143–9.

192. Martin WS, Duncan WJ, Powers TL, Sawyer JC. Costs and benefits of selected response inducement techniques in mail survey research. *Journal of Business Research* 1989;**19**:67–79.

193. Faria AJ, Dickinson JR, Filipic TV. The effect of telephone versus letter prenotification on mail survey response rate, speed, quality and cost. *Journal of the Market Research Society* 1990;**32**:551–68.

194. Czaja R, Blair J. Designing surveys: a guide to decisions and procedures. Thousand Oaks, CA: Pine Forge Press; 1995.

195. Nevin JR, Ford NM. Effects of a deadline and a veiled threat on mail survey responses. *Journal of Applied Psychology* 1976;**61**:116–18.

196. Kahle LR, Sales BD. Personalization of the outside envelope in mail surveys. *Public Opinion Quarterly* 1978;**42**:547–50.

197. Swan JE, Epley DE, Burns WL. Can follow-up response rates to a mail survey be increased by including another copy of the questionnaire? *Psychological Reports* 1980;**47**:103–6.

198. Blass T, Leichtman SR, Brown RA. The effect of perceived consensus and implied threat upon responses to mail surveys. *Journal of Social Psychology* 1981;**113**:213–16.

199. Dommeyer CJ. The effects of negative cover letter appeals on mail survey response. *Journal of the Market Research Society* 1987;**29**:445–51.

200. Gitelson RJ, Drogin EB. An experiment on the efficacy of a certified final mailing. *Journal of Leisure Research* 1992;**24**:72–8.

201. Roberts H, Pearson JCG, Dengler R. Impact of a postcard versus a questionnaire as a first reminder in a postal lifestyle survey. *Journal of Epidemiology and Community Health* 1993;**47**:334–5.

202. Duncan W. Mail questionnaires in survey research: a review of respondent inducement techniques. *Journal of Management* 1979;**5**:39–55.

203. Armstrong JS, Lusk EJ. Return postage in mail surveys: a meta-analysis. *Public Opinion Quarterly* 1987;**51**:233–48.

204. Harris JR, Guffey HJ. Questionnaire returns: stamps versus business reply envelopes revisited. *Journal of Marketing Research* 1978;**15**:290–3.

205. Jones WH, Linda G. Multiple criteria effects in a mail survey experiment. *Journal of Marketing Research* 1978;**15**:280–4.

206. Labrecque DP. A response rate experiment using mail questionnaires. *Journal of Marketing* 1978;**42**:82–3.

207. Hopkins KD, Podolak J. Class-of-mail and the effects of monetary gratuity on the response rates of mailed questionnaires. *Journal of Experimental Education* 1983;**51**:169–70.

208. Corcoran KJ. Enhancing the response rate in survey research. *Social Work Research and Abstracts* 1985;**21**:2.

209. Elkind M, Tryon GS, de Vito AJ. Effects of type of postage and covering envelope on response rates in a mail survey. *Psychological Reports* 1986;**59**:279–83.

210. Harvey L. A research note on the impact of class-of-mail on response rates to mailed questionnaires. *Journal of the Market Research Society* 1986;**28**:299–300.

211. Cartwright A, Windsor J. Some further experiments with factors that might affect the response to postal questionnaires. *Survey Methodology Bulletin* 1989;**25**:11–15.

212. Zeinio RN. Data collection techniques: mail questionnaires. *American Journal of Hospital Pharmacy* 1980;**37**:1113–19.

213. Jones WH. Generalizing mail survey inducement methods: population interactions with anonymity and sponsorship. *Public Opinion Quarterly* 1979;**43**:102–11.

214. McDaniel SW, Rao CP. An investigation of respondent anonymity's effect on mailed questionnaire response rate and quality. *Journal of the Market Research Society* 1981;**23**:150–60.

215. Campbell MJ, Waters WE. Does anonymity increase response rate in postal questionnaire surveys about sensitive subjects? A randomised trial. *Journal of Epidemiology and Community Health* 1990;**44**:75–6.

216. McKee DO. The effect of using a questionnaire identification code and message about non-response follow-up plans on mail survey response characteristics. *Journal of the Market Research Society* 1992;**34**:179–91.

217. Wiseman F. A reassessment of the effects of personalization on response patterns in mail surveys. *Journal of Marketing Research* 1976;**13**:110–11.

218. Andreasen A. Personalizing mail questionnaire correspondence. *Public Opinion Quarterly* 1970;**34**:273–7.

219. Houston MJ, Jefferson RW. The negative effects of personalisation on response patterns in mail surveys. *Journal of Marketing Research* 1975;**11**:413–17.

220. Trice AD. Elements of personalization in covering letters may affect response rates in mail surveys: a further analysis of Worthen and Valcarce (1985). *Psychological Reports* 1986;**58**:82.

221. Roberts RE, McCrory OF, Forthofer RN. Further evidence on using a deadline to stimulate responses to a mail survey. *Public Opinion Quarterly* 1978;**42**:407–10.

222. Childers TL, Pride WM, Ferrell OC. A reassessment of the effects of appeals on response to mail surveys. *Journal of Marketing Research* 1980;**17**:365–70.

223. Woodward JM, McKelvie SJ. Effects of topical interest and mode of address on response to mail survey. *Psychological Reports* 1985;**57**:929–30

224. Worthen BR, Valcarce RW. Relative effectiveness of personalized and form covering letters in initial and follow-up mail surveys. *Psychological Reports* 1985;**57**:735–44

225. Green KE, Stager SF. The effects of personalization, sex, locale, and level taught on educators' responses to a mail survey. American Educational Research Association Annual Meeting (1986, San Francisco, California). *Journal of Experimental Education* 1986;**54**:203–6.

226. Wunder GC, Wynn GW. The effects of address personalisation on mailed questionnaires' response rate, time and quality. *Journal of the Market Research Society* 1988;**30**:95–101.

227. Green KE, Kvidahl RF. Personalization and offers of results: effects on response rates. *Journal of Experimental Education* 1989;**57**:263–70.

228. Salomone PR, Miller GC. Increasing the response rates of rehabilitation counselors to mailed questionnaires. *Rehabilitation Counseling Bulletin* 1978;**22**:138–41.

229. McKillip J, Lockhart DC. The effectiveness of cover-letter appeals. *Journal of Social Psychology* 1984;**122**:85–91.

230. Wagner WG, O'Toole WM. The effects of cover letter format on faculty response rate in mail survey research. *Educational and Psychological Research* 1985;**5**:29–37.

231. Dodd DK, Markwiese BJ. Survey response rate as a function of personalized signature on cover letter. *Journal of Social Psychology* 1987;**127**:97–8.

232. Biner PM. Effects of cover letter appeal and monetary incentives on survey response: a reactance theory application. *Basic and Applied Social Psychology* 1988;**9**:99–106.

233. Dodd DK, Boswell DL, Litwin WJ. Survey response rate as a function of number of signatures, signature ink color, and postscript on covering letter. *Psychological Reports* 1988;**63**:538.

234. Biner PM, Barton DL. Justifying the enclosure of monetary incentives in mail survey cover letters. *Psychology and Marketing* 1990;**7**:153–62.

235. Katz S. The functional approach to attitude change. *Public Opinion Quarterly* 1960;**24**:163–204.

236. Brehm JW. A theory of psychological reactance. New York: Academic Press; 1966.

237. Wicklund RA. Freedom and reactance. Potomac, MT: Lawrence Erlbaum; 1974.

238. Brehm SS, Brehm JW. Psychological reactance: a theory of freedom and control. New York: Academic Press; 1981.

239. Smith WC, Crombie IK, Campion PD, Knox JD. Comparison of response rates to a postal questionnaire from a general practice and a research unit. *British Medical Journal Clinical Research Edition* 1985;**291**:1483–5.

240. Dommeyer CJ. Does response to an offer of mail survey results interact with questionnaire interest? *Journal of the Market Research Society* 1985;**27**:27–38.

241. Hansen RA. A self-perception interpretation of the effect of monetary and nonmonetary incentives on mail survey respondent behavior. *Journal of Marketing Research* 1980;**17**:77–83.

242. Bem DJ. Self perception theory. In: Berkowitz L, editor. Advances in experimental social psychology. New York: Academic Press;1972. p. 1–62.

243. Furse DH, Stewart DW. Cognitive dissonance and response to mail surveys: working paper no: 81-119. Nashville, TN: Vanderbilt University; 1981.

244. Furse DH, Stewart DW. Monetary incentives versus promised contribution to charity: new evidence on mail survey response. *Journal of Marketing Research* 1982;**19**:375–80.

245. Adams JS. Inequity in social exchange. In: Berkowitz L, editor. Advances in experimental social psychology. New York: Academic Press; 1965. p. 267–99.

246. Berry SH, Kanouse DE. Physician response to a mailed survey: an experiment in timing of payment. *Public Opinion Quarterly* 1987;**51**:102–14.

247. Armstrong JS. Monetary incentives in mail surveys. *Public Opinion Quarterly* 1975;**39**:111–16.

248. Hopkins KD, Gullickson AR. Response rates in survey research: a meta-analysis of the effects of monetary gratuities. *Journal of Experimental Education* 1992;**61**:52–62.

249. Whitmore WJ. Mail survey premiums and response bias. *Journal of Marketing Research* 1976;**13**:46–50.

250. Little RE, Davis AK. Effectiveness of various methods of contact and reimbursement on response rates of pregnant women to a mail questionnaire. *American Journal of Epidemiology* 1984;**120**:161–3.

251. Paolillo JG, Lorenzi P. Monetary incentives and mail questionnaire response rates. *Journal of Advertising* 1984;**13**:46–8.

252. Cook JR, Schoeps N, Kim S. Program responses to mail surveys as a function of monetary incentives. *Psychological Reports* 1985;**57**:366.

253. Mortagy AK, Howell JB, Waters WE. A useless raffle. *Journal of Epidemiology and Community Health* 1985;**39**:183–4.

254. Woodward A, Douglas B, Miles H. Chance of free dinner increases response to mail questionnaire [letter]. *International Journal of Epidemiology* 1985;**14**:641–2.

255. Blythe BJ. Increasing mailed survey responses with a lottery. *Social Work Research and Abstracts* 1986;**22**:18–19.

256. Weltzien RT, McIntyre TJ, Ernst JA, Walsh JA. Crossvalidation of some psychometric properties of the CSQ and its differential return rate as a function of token financial incentive. *Community Mental Health Journal* 1986;**22**:49–55.

257. Dommeyer CJ. How form of the monetary incentive affects mail survey response. *Journal of the Market Research Society* 1988;**30**:379–85.

258. Hubbard R, Little EL. Promised contributions to charity and mail survey responses: replication with extension. *Public Opinion Quarterly* 1988;**52**:223–30.

259. Brennan M. The effect of a monetary incentive on mail survey response rates: new data. *Journal of the Market Research Society* 1992;**34**:173–7.

260. Mosher DL. Measurement of guilt in females by self-report inventories. *Journal of Consulting and Clinical Psychology* 1968;**32**:690–5.

261. Erdos PL, Morgan AJ. Professional mail surveys. New York: McGraw-Hill; 1970.

262. Dommeyer CJ. Offering mail survey results in a lift letter. *Journal of the Market Research Society* 1989;**31**:399–408.

263. Lovelock CH, Stiff R, Cullwick D, Kaufman IM. An evaluation of the effectiveness of drop-off questionnaire delivery. *Journal of Marketing Research* 1976;**13**:358–64.

264. Salvesen KA, Vatten LJ. Effect of a newspaper article on the response to a postal questionnaire. *Journal of Epidemiology and Community Health* 1992;**46**:86.

265. National Centre for Social Research, Picker Europe, Department of Primary Health Care and General Practice. National surveys of NHS patients: general practice 1998. London: NHS Executive; 1999.

266. Erens B, Primatesta P, Prior G. Health survey for England: the health of minority ethnic groups 1999. London: National Statistics; 1999.

Appendix 1

Guidance on other aspects of survey design and administration

In the review proper, those aspects of the survey process that were not amenable to experimental manipulation were excluded. For completeness, this appendix provides some general guidance on those excluded aspects, based primarily on the accumulated experience of the authors, with selected references to sources of expert opinion.

Sampling for survey research

Sampling of survey respondents is efficient, saving time and resources that would be required for a census of all population members (for many populations a complete census would in any case be impracticable). Statistical techniques can be applied to data that have been collected by using scientifically selected samples, to derive parameter estimates (e.g. of the prevalence and incidence of attributes, attitudes and behaviour, and the distributions of ages and scores on attitude scales) in the population from which the sample is drawn.

However, despite these important advantages, all sampling introduces random sampling variation. The extent of this needs to be taken into account in both designing samples and interpreting results.

Definition of an adequate sample

An adequate sample is one that:

- is selected by using an unbiased method (a biased method is one that produces results that will differ from the true population values in a consistent or systematic way)
- is representative of the underlying population of interest
- is sufficiently large to make inferences about the underlying population within acceptable margins of random variability (i.e. with sufficiently narrow CIs).

To achieve an adequate sample, the survey researcher needs to:

- state the objectives of the survey clearly and precisely
- define explicitly the population to be surveyed, in terms of inclusion and exclusion criteria
- choose a sampling frame that is appropriate to the defined study population
- specify rigorous and objective sample selection methods (preferably probability sampling methods)
- determine the required achieved sample size, taking into account the likely variation in the characteristics of interest in the population, the size of differences between subgroups that the researcher wishes to be able to detect, and the level of confidence required in estimates of population values derived from the sample
- contact more than the required sample to allow for the losses that are expected due to the occurrence of ineligibles (e.g. those who do not fit population criteria; those who have died) and to non-response (e.g. non-contacts; refusals).

Some sampling terms

In the literature on sampling,[1–4] a number of common words and terms are used in a particular way. The term "population" is used to denote the complete set of units from which a sample is selected and to which the sample-based results will apply; these "units" or "elements" could be patients, or members of the public, or hospitals, but they could also be events such as births or attendances at a clinic. It is important to recognise that there may be different sampling units at different stages in the selection (e.g. multistage sampling where the initial sample may be of general practices, followed by a second-stage sampling of patients within practices). Note also that the sampling units may or may not be identical with survey respondents (e.g. the sampling unit could be the seat in the waiting room, but the survey respondent would be the patient occupying that seat). Regardless of these considerations, however, a precise operational definition of population and units is crucial (i.e. there must be explicit eligibility criteria). The definition of the target population should relate explicitly to the research aims. It should specify explicit inclusion criteria (e.g. "adult patients attending the diabetic clinic at 'X' hospital") and exclusion criteria (e.g. "those attending the clinic for a second or subsequent time during the sampling period and out-of-area

cases") and may also involve a time-frame (e.g. "those attending during the months of June and July"). Of course, when a survey is being used to collect data within the framework of a trial – as will be the case in many health technology assessments – the population will probably have already been defined for the "parent" trial.

Another commonly used term is "sampling frame", which is a listing of all the units (elements) in the population that are eligible to be sampled. Ideally, the sampling frame should correspond exactly to the target population. In practice, however, the only available sampling frame used may itself be a sample (adequate or inadequate) of the real target population (e.g. individuals on the electoral register as a substitute for all adults living in a particular area). Moreover, populations and sampling frames may be implicit because no physical listing exists (e.g. patients attending a sexually-transmitted disease clinic who are considered as a sample from the population of all who have attended the clinic in the recent past and (barring major changes) will attend in the immediate future).

The ideal sampling frame is a listing of population units that:[5]

- matches the target population one-to-one
- is comprehensive (has no omissions)
- has one entry only for each eligible unit
- contains no ineligible units
- contains complete, accurate and up-to-date identifying and tracing information for each unit
- contains information about each unit that is useful for sample stratification (e.g. age, sex)
- is accessible and cleared for use in research
- ideally, can be manipulated by computer.

In practice, few sampling frames – including many of those commonly used in health surveys, such as the electoral register, the Postcode Address File, and general practitioner age, sex and morbidity registers – exactly satisfy all these criteria. Because the sampling frame is so fundamental to the enquiry and because defects can affect the validity of the sample, it is often worth spending time and effort on improving it.

Avoiding sampling bias

Bias can be introduced in the sampling process. One source of such bias is where the sampling frame is not an adequate representation of the underlying population. For example, electoral registers are typically biased against students and other "floating citizens" who have not registered to vote, or who are ineligible to vote (e.g. the homeless, foreign nationals). Selection error may arise if a non-probability method of sampling (i.e. a method that does not give each member of the underlying population a known chance of being included) is used. For example, "invited" samples, such as reader surveys carried out by a journal, are typically non-representative; the readership of even a professional journal is unlikely to be truly representative of all members of that profession.

To minimise sampling bias, it is important to ask:

- To what population (e.g. professionals, patients, institutions) are the results intended to apply?
- Do the sampling frame and selection procedure used to select the sample give all members of the target population a known chance of selection? If not, how many and what types of population units are likely to be excluded?
- Is it possible to ensure, or reasonably assume, that the units to which there is access are a random sample of all target population units (or is there bias with respect to those that are accessible)?

If the answer to any of the preceding three questions is "no", there is a sampling frame bias.

It is also important to query:

- Do all members of the population have a known probability of selection (e.g. an equal chance of being chosen)?
- Are some population units listed or available for selection several times? For example, this could occur if some individuals appear more than once on the population listing. When the sampling units are events rather than individuals, it would be an issue if some individuals gave rise to more than one event (e.g. visiting a doctor's surgery several times during the sampling period). It is important to note that, if some population units are listed more than once, there is still bias even if they are not selected more than once (or at all).

All statistical inference from a sample to the underlying population and all hypothesis testing assume the random selection of the sampled units. It is therefore important to clarify whether the chosen selection process is truly random. It could be non-random because of: concessions to convenience (e.g. going for the "easy-to-find" cases – in an interview survey, those people who are at home during daylight hours); a wish to include "interesting" cases; or a desire to exclude difficult

or uninteresting cases. In particular, it is important to recognise that human beings cannot choose a random sample by judgement alone; there has to be a random selection procedure.

Secondly, the most perfectly constructed sample can be largely invalidated for the purposes of drawing conclusions and making inferences if there is gross and differential non-response at the data collection stage (i.e. non-response bias). If sample members of a particular kind are less likely than average to respond, this is equivalent to under-sampling them (and relatively over-sampling other groups) in an uncontrolled way.[5]

Sample selection methods

As noted above, statistical inference is predicated on the assumption that probability (random) sampling methods have been used in selecting respondents. For probability sampling, each unit in the target population must have a calculable, non-zero probability of being selected. Probability sampling techniques include:

- Simple random sampling: This method generally uses a paper-based random numbers table or a computerised random selection procedure. Each selection is made independently and each unit has an equal probability of being selected.
- Systematic random sampling: This uses a random start in the population listing, then selects every *n*th unit. It has the merit of being easier to implement, especially by hand, but it may, however, lead to selection bias if the list is organised in some systematic manner. For example, in sampling nursing staff from a series of wards, where each ward has around 20 staff and the staff are listed in order of seniority within a ward, a low random start and a sampling interval of ten will tend always to select one very senior staff member and one from the middle of the seniority list and to under-sample the remainder. Such risks can generally be removed through paying attention to how the list was compiled
- Stratified random sampling: This involves controlling the composition of the sample in relevant respect(s), such as age and/or gender. Stratification can be applied at any stage of selection (but only, of course, if information about all population units is available). Prior to sampling, units are split into subsets or strata that are likely to differ in terms of what the survey aims to measure and separate random samples are then drawn within each stratum. For example, if the population contains men and women (who can be pre-identified) and if the results for men are likely to differ from the results for women, then something is gained by listing and sampling the two sexes as separate strata, to predetermine the number of men and the number of women selected. Provided the probabilities of selection are known, estimates from strata can be combined to give estimates for the total population.[2,3] Stratification can significantly improve the precision of estimates, but its effect is, in most applications, small relative to that of sample size (see below).
- Multistage (clustered) random sampling: This is carried out in stages, using different population units at each stage. For example, in a two-stage random sampling procedure, the first stage units (e.g. areas, hospitals) are randomly selected from an appropriate sampling frame at stage one, and the second (final) stage units (patients) are selected at stage two from within selected stage one units. Thus the final stage units are said to be clustered within first stage units. Multistage sampling reduces the amount of administrative, sample selection and data collection effort required by the researchers. The results of multistage sampling will be unbiased, but they are usually less precise (i.e. have wider CIs) than those of single-stage sampling.

Calculating sample size

If units have been selected at random, the precision of sample-based population estimates is then mainly determined by sample size: the larger the sample size, the smaller the random sampling error. With large populations (e.g. all adult residents of a health authority catchment area), what matters is the absolute size of the sample. In such large populations, the proportion of the population included in the sample is unimportant. Hence, despite the inequality in size of the two countries, to provide results of the same level of precision for the population of Scotland and the population of England, equal-sized samples are generally appropriate. It is only when the number of units to be selected for the sample exceeds about 10% of the units in the population that the finite population correction factor needs to be considered in calculating the precision of results.[2,3]

Sample precision increases approximately in proportion to the square root of the sample size. Increasing sample size by a factor of ten therefore increases precision (narrows CIs) by only just over threefold.

In research studies in which the survey is embedded within a particular study design (e.g. a randomised controlled trial), the sample size will generally be based on a calculation of the statistical power

required to test specified hypotheses. If a survey is being carried out as a separate operation, sample size calculations should be based on the desired precision of the most important estimates to be made from study findings. Fowler[4] highlights the importance of considering multiple "outcome measures" in calculating sample sizes because the precision needed for different estimates is likely to vary.

Estimation of the required sample size for major or complex studies is fundamental and requires a knowledge of sampling statistics, so a trained statistician should be consulted if possible. The statistician will want to know:

- What is the null hypothesis (if any)? (e.g. "no difference in the prevalence of symptoms between men and women")
- How confident do you want to be in accepting or rejecting the null hypothesis? (usually 95% or 99%)
- How certain do you want to be of detecting, for example, a real (population) difference between men and women of, say, 5% in symptom prevalence? (usually 80% or 90%)
- If you are not testing a hypothesis but simply want to make a point and interval estimate (e.g. of the overall symptom prevalence) in the underlying population, how precise do you want that estimate to be? (e.g. within ±3%)
- How variable is what you are trying to estimate (e.g. situation at a point in time; change over time) likely to be across the population?
  - for a dichotomous variable, the proportion with that attribute (e.g. proportion of the population, or change in the proportion of the population, exhibiting a particular symptom)
  - for a continuous variable, the population standard deviation of what is being measured (e.g. diastolic blood pressure).

This requirement of having some preconceived idea about variability means that, in order to calculate a sample size, the researcher and the statistical adviser will have to "guesstimate" some of the very things the researcher hoped to measure! Fortunately, it is usually possible to make the requisite guesses to a sufficient degree of precision (e.g. on the basis of published data, previous research, or pilot work).[5]

Finally, it should be remembered that sample size calculations refer to achieved sample sizes. As 100% response rates are rarely, if ever, achieved, the survey researcher needs to estimate likely response rates and to over-sample accordingly. For example, if the target sample size is 500 and the anticipated questionnaire completion and return rate is 70%, approximately 714 individuals must be sampled and contacted to attain the desired sample size.

Sources of survey questions

Many surveys on health topics will address issues and concepts that have been the subject of previous research and surveys, and it is important to avoid "re-inventing the wheel". Apart from the advantages of drawing upon the expertise and experience of others, the development and refining of new questions, ensuring that they are valid and reliable, is time-consuming and expensive. In many circumstances, use can be made of existing well-validated questions or even whole questionnaires. Of course, it is important to recognise that the existence in the literature of a set of questions that has a relevant-sounding label does not guarantee that it is appropriate to a particular population and study. It is also important to note that questions and scales that have previously been developed and applied in one setting (e.g. secondary care in the USA) may not be readily transferable to a different setting (e.g. primary care in the UK). At a minimum, it may be necessary to test the validity and reliability of the questions in the new setting; in some cases, an extensive exercise in cross-cultural adaptation may be required.[6,7]

Questions on health status and quality of life

Many surveys conducted in the area of public health, or in the context of health technology assessment, have as key outcome variables the "health status" or "quality of life" of respondents. There is a lack of consensus on what either of these terms actually mean and neither is easy to measure in ways that meet criteria of validity, reliability, sensitivity, responsiveness to change and so on.

There is nevertheless a wealth of instruments, with established validity and reliability, for measuring health status, both in general populations and in specific disease- and age-groups. Rather than trying to develop and test a new set of questions, the researcher should first review these existing instruments and see whether any are appropriate to the aims and objectives of the planned survey. Comprehensive reviews of instruments for measuring health status and quality of life, including details of reported validity, reliability and responsiveness to change, are available.[8–11] Of course, no instrument should be chosen unthinkingly; candidate scales need to be evaluated against explicit criteria to

ensure that they are appropriate to the survey's aims and objectives.[7,12]

The Centre for Applied Social Surveys Question Bank

This Question Bank is a resource designed to aid questionnaire developers who are in search of appropriate questions for health and other surveys. It is a website that is maintained and continually updated by a team at the Centre for Applied Social Surveys, a resource centre supported by the Economic and Social Research Council. The Question Bank is a store of questionnaires from important and established surveys, reproduced in their original format. In addition to questionnaires, it contains commentary on concepts and measurement issues, through which the wording, context, origin, purpose and performance of questions can be better understood. However, it does not contain datasets or substantive survey reports. The site search engine enables users to scan questionnaires and to locate questions and text containing particular words or phrases. The Question Bank contains questionnaires from a number of important surveys that cover health topics, including successive annual versions of the Health Survey for England and the General Household Survey, and also the National Patients Survey, the Welsh Health Survey, the National Survey of Sexual Attitudes and Lifestyles, and others. The address of the site is: http://qb.soc.survey.ac.uk/

Piloting and pretesting questionnaires and survey procedures

Unless a questionnaire consists entirely of previously tested and validated questions that have been successfully used together before, it is advisable to go through a process of pretesting and piloting. This procedure helps to ensure face and content validity, and may indicate the need for rewording, reformatting or other refinements to the questionnaire itself or to the proposed conduct of the survey. Provision should be made in the budget and timetable for planning and conducting pretests and pilot studies, and interpreting, feeding back and acting upon the results.

There are broadly three types of pretest or pilot study, each with a different purpose:

- developmental trials and experiments:
  - to explore new topic areas
  - to test the feasibility of novel methods
  - to develop the survey instruments
  - to focus on particular problem areas
- cognitive tests of questions and instruments:
  - to check how respondents cope with the questionnaire and instructions
  - to check whether respondents understand the questions as intended
  - to check how the respondents set about answering the questions
- rehearsal pilots:
  - to test the survey procedures overall under "main survey" conditions
  - to detect and remove minor "glitches"
  - to provide better estimates on which to base sample size calculations
  - to estimate the rate of, and speed of, response to the main survey
  - to check timings
  - to smooth co-ordination and establish systems and routines.

Priorities for pretest and pilot work

For an important survey in an area that has not been previously explored, all three types of pretest and pilot work may be required. Exploratory work, often using qualitative methods such as in-depth interviews, can be fundamental in identifying topics and content area. In situations when the basic topic and approach are clear, the priority will be cognitive testing and rehearsal pilots. These two approaches do different jobs and are not interchangeable.
If it seems that administration of the survey will not be problematical, then the priority should be the cognitive testing of questions, questionnaires and other survey documents (e.g. covering letter and any supplementary information; reminder letters), especially when the questionnaire contains questions or instructions not previously tested in the population of interest. Cognitive testing is particularly important in self-completion questionnaires because there is no opportunity for feedback to interviewers on how the respondents "took to" and appeared to understand the questions, and there is no possibility of skilled interviewing making up for poor questionnaire design.

Cognitive pretesting

Cognitive pretesting methods[13–16] are about uncovering what goes on in the minds of respondents when they receive the "package" sent to them, which is likely to consist of a covering letter, the questionnaire and a return envelope.

A weak link in the process of conducting postal or other self-completion surveys is the task of persuading potential respondents to complete and return the questionnaire, to yield a high response rate. If respondents are to return the questionnaire, they must: open the package, read the

introductory message (in the covering letter), start filling in the questionnaire, understand the questions, answer the questions completely and in the way intended by the researcher, and return the completed questionnaire. Many factors can affect the chance that this will happen.[17,18] Cognitive testing can give clues to how well different aspects of the postal package (e.g. outgoing and return envelopes; stationery and letterhead; size and weight of the package; look and presentation of the contents) are working: whether they attract the interest and curiosity of the respondent, or act as a turn-off. Cognitive testing can also indicate if the covering letter is fulfilling the aims of communicating with the respondent regarding: the sponsorship and aims of the study; the way in which sampled individuals have been selected; the importance of and implications of participating; how data will be handled; how confidentiality will be maintained; how and in what form the results will be made available; and how respondents can make contact with those conducting the survey.

Comprehension of questions may also be a problem, particularly if the survey researcher has little experience of communicating with the types of people who it is hoped will respond. Inexperienced designers may also misjudge if respondents will be willing and able to give adequate answers. Cognitive testing can indicate how respondents are interpreting questions, response categories and instructions, and how they are going about formulating their answers.

The main cognitive testing techniques require direct interaction between a researcher, or specially trained interviewers, and the test respondents, who should be drawn from the same population as the potential respondents for the main survey. There is no hard and fast rule about sample size; samples are usually relatively small (not more than 20). However, in contrast to the random, probability sampling procedures to be employed in the main survey, the sample is purposively selected to ensure the representation of groups who may have different problems with responding (e.g. to ensure a good spread of age groups, levels of education and literacy). The tests can take place in the respon-dents' own homes or at a central location. A small fee and/or travel expenses are usually paid to respondents because a significant effort is required of them.

Each test respondent is asked to go through a procedure that mimics as closely as possible what self-completion respondents will be required to do: open the package; read the covering letter; fill in the questionnaire; and place the completed questionnaire in the return envelope. If desired, the procedure can be confined to completing the questionnaire or key parts thereof. Test respondents provide feedback to the researcher or interviewer, which is then summarised into action points for revising the survey documents and procedures.

One method of providing this feedback is the "think aloud" approach, in which respondents are asked to verbalise their reactions and thought processes as they go through the required steps, while the researcher or interviewer takes notes or tape-records these comments. However, not all test respondents can cope with verbalising their thought processes as they go along. An alternative approach is the "debriefing interview", in which the test respondent completes the required steps in self-completion mode, unobtrusively observed by the researcher or interviewer (who may take notes). Respondents are then taken through the steps again, with the interviewer prompting with probes such as the following:

- What did you think when you first opened the package?
- What did the letter tell you?
- Did you happen to notice this instruction before the question?
- What did you think you needed to do to answer that question?
- What sorts of things were you thinking about when you answered that question?
- I noticed that you hesitated before you answered that question. Why was that?

Once again, the interviewer takes notes or tape-records the responses to these probing questions.

Sometimes there will not be the time or resources to carry out cognitive testing. A cheaper and quicker alternative is review by an expert panel. This is a technique that can be used when the survey researchers have access to colleagues who are experienced in survey work, but who are not directly involved in the current survey. The proposed documents and procedures are sent to panel members for review. Written comments may be given or an informal seminar may be arranged, at which the designers and panel members exchange views and suggestions for improvement. Expert panels may also be used as a preliminary to cognitive testing.

At the very minimum, even if none of the formal procedures described above is used,

the questionnaire and any accompanying documentation should be shown to colleagues, family and/or friends; a fresh eye can often pick out ambiguous or confusing questions or instructions.

After whatever pretesting approach has been used, the questionnaire should be modified (if necessary) to take account of the feedback. If extensive modifications need to be made, it may be advisable to repeat the pretest.

Rehearsal pilots

Once a final draft questionnaire is available, it should be pilot tested, ideally with a larger sample (often 30–100, although the actual sample size will generally be dictated by resources). The sample used should again represent the variation in the types of respondent and respondent circumstances that will be met with in the main survey. In the pilot test it is desirable to match methods of administration to real life circumstances (e.g. if the questionnaire is intended for self-completion in a waiting room, the pilot test should be carried out under those conditions). The aims of the rehearsal pilot study are:

- to look for how well questions work; poor questions may be indicated by:
  - frequently omitted questions
  - inappropriate responses
  - inconsistent responses
  - lack of spread of responses
  - an apparent need for new response categories (use of open-ended questions may be appropriate at this stage, to inform the development of response categories for closed questions for the survey proper)
- to estimate likely response rates
- to identify likely non-response bias.

Once again, the questionnaire and data collection procedures should be refined on the basis of the findings from the pilot test. As with the pretest phase, if extensive changes are required, it may be necessary to repeat the pilot test.

Finally, when all the changes to the questionnaire have been made, it should be carefully proofread before being printed and distributed.

Assessing validity and reliability in surveys

Validity – whether a question and its associated response options are actually measuring what they purport to measure – is a key issue in survey research. In designing questions, or testing questions previously used in one context or setting in a new population, the survey researcher needs to pay careful attention to whether the information yielded is valid. One aspect of validity is "criterion validity": if the question/questionnaire yields results that correspond with those obtained by another, "gold-standard" method (ideally an objective measure) applied simultaneously ("concurrent validity"), or that forecast a criterion value ("predictive validity"). For example, declared smoking status could be validated by a measurement of cotinine in the saliva. Criterion validity is generally assessed formally by using statistical techniques such as correlation. A major problem, however, with assessing criterion validity is a lack of appropriate gold-standard measures. In interviewer-administered surveys, it may be possible to validate responses to some questions by direct observation. Documentary evidence may also be available as a source of validation.

Another type of validity is "face validity", whether, "on the face of it", the questions are measuring what they are supposed to measure. This is generally assessed informally, by asking non-expert and untrained "judges" (for example, colleagues, family or friends) to examine the questionnaire to see whether the items look satisfactory to them. However, face validity alone is not a sufficient test of the validity of the questions.

Similar to face validity is "content validity", which concerns whether the choice of items and the relative importance given to each are appropriate in the eyes of those who have some knowledge of the topic area. This is best achieved by having the questionnaire critiqued by a panel of people who are knowledgeable about the topic, including members of the target population. This critique involves assessing if the questionnaire covers everything it should and does not include extraneous matter. However, favourable assessments of content validity do not in themselves guarantee that the measure will produce valid information.

"Construct validity" refers to whether the results obtained using the questionnaire confirm expected statistical relationships, the expectations being derived from underlying theory. In drafting questions, theoretical assumptions are always made about how concepts are related to one another; these assumptions should be tested. One of the most common ways of assessing construct validity is through a test of known-group validity.[19] This involves making comparisons across groups who would *a priori* (either on the basis of theory, or by

drawing on previous empirical evidence) be expected to yield different results. For example, if a questionnaire were designed to assess health status, one would expect people with diagnosed, chronic disease to have poorer scores than those with no known current illnesses. Similarly, because trends over a number of years have shown an inverse relationship between social class and smoking behaviour, one would expect to observe higher rates of smoking in respondents of lower socio-economic status. Another way of establishing construct validity is through multitrait, multi-method analysis.[20] This is most appropriate in the assessment of the validity of scales designed to measure some state or trait (e.g. health status, job satisfaction). It involves administering more than one instrument purporting to measure similar and dissimilar traits (e.g. in relation to health status, the SF-36 and the Nottingham Health Profile) and examining the correlation between scores on the various instruments. Higher correlations between scores on domains that measure similar concepts (e.g. physical function and role limitation due to physical impairments), either within one instrument or across instruments, and weaker correlations between domains measuring dissimilar traits (e.g. physical function and mental health) are indicative of construct validity.

Finally, it may be possible to assess "freedom from absolute or relative bias", which indicates if the question/questionnaire yields results that fairly reflect the distribution of some target variable in the population and in subpopulations. An example of absolute bias would be if a question, or a set of questions, that may be valid in some senses as a measure of disability, is still open to objection because it gives too high or too low an estimate of the prevalence of disability or a distorted distribution of the severity of disability in the population. An example of relative bias would be when the questions obtained responses from elderly people that made them seem less disabled than younger people with similar objective incapacities.

It is also important that questions and survey procedures should yield reliable or reproducible findings. As with validity, there are a number of different approaches to measuring reliability.

In the case of response scales (where responses to a number of similar questions designed to measure the same construct are combined to yield an overall score), reliability may be assessed through formal statistical measures of homogeneity[19] or "internal consistency". The idea here is that all questions suffer from some degree of response unreliability, but that the degree of logical and conceptual consistency found between responses to questions designed to capture the same property of a subject (e.g. satisfaction with health care provision) provides an indication of the reliability of those responses. One appropriate statistic is the item-total correlation,[19] which measures the strength of the correlation between each item and its constituent scale, with a rule of thumb being that the absolute value of the correlation coefficient should be at least 0.2. Another is Cronbach's alpha,[21] which, as Fitzpatrick and colleagues explain (p. 23), "essentially estimates the average level of agreement of all possible ways of performing split-half tests".[12] The higher the value, the greater the internal consistency, with criterion values for adequate reliability lying between 0.7 and 0.9.[22] Internal consistency for questions designed to address similar concepts, but not to be combined into an overall scale score, can similarly be assessed through correlation measures such as intra-class correlations.[19]

"Test-retest reliability" is the most logically straightforward measure of reliability. It involves checking whether the same answer is obtained if the question is asked of the same individual at two points in time, during which period no real change has occurred in that individual in relevant respects. It is important to choose an interval between the two measurements that is long enough so that respondents are not simply recalling and repeating their initial answer, but is not so long that real change may have taken place. In practice it is often difficult to apply a satisfactory test-retest check because of the difficulty of simultaneously satisfying both the "no recall bias" and the "no real change" conditions.

Survey management

Adequate sampling, good questionnaire and survey design, and good piloting are not enough; the whole survey process needs to be carefully managed. The survey researcher must: estimate and procure materials, personnel and other resources; timetable survey activities and monitor progress against the defined schedule; manage the dispatch and return of questionnaires; take appropriate measures to ensure good response rates; monitor data quality; and oversee data management. Surprisingly, perhaps, these aspects of the survey process are amongst the least well researched and reported, with most survey experts (the authors of this review included) relying on

tried and proven methods. Below, are presented recommendations for good practice in respect of survey manage-ment, with particular, although not exclusive, emphasis on postal and other self-completion surveys. (Further guidance on the administration and management of telephone and face-to-face interviews is provided by Morton Williams,[23] by Dillman,[17] and by Salant and Dillman,[24] among others.[25,26])

Determining survey resource needs

Fink[26] has suggested that the following questions should be asked in determining survey resource needs:

- What are the main tasks to be carried out?
- What skills are needed to carry out these tasks?
- How much time is needed for each task?
- Who can be used to perform each task?
- What are the costs of each task?
- What additional resources are needed?

Human resource functions in questionnaire surveys

Human resource functions and skills may be divided into three broad categories:

- research functions:
  - designing questions
  - designing questionnaires (including layout)
  - piloting questionnaires
  - designing protocols for recruitment and retention of respondents
  - sampling respondents
  - checking and coding data
  - analysing and interpreting data
  - reporting findings
- field data collection functions (more applicable to interviewer-administered and to "captive audience" self-completion surveys than to postal surveys):
  - approaching sampled individuals and securing their co-operation
  - interviewing respondents
  - coding data in the field
- administrative and clerical functions:
  - word-processing questionnaires
  - database set-up and management
  - stuffing and opening envelopes
  - handling enquiries from respondents.

Other resources required

Other resources that must be budgeted for and procured include:

- printing/copying facilities:
  - for questionnaires
  - for covering letters
  - for other survey documents
- postage facilities (stamps, business reply envelopes or franking facilities):
  - for outgoing mail
  - for returned questionnaires
- telephone facilities (including perhaps answering machine or voice-mail facilities):
  - for outgoing calls (especially when arrangements to interview are being made by telephone, or telephone follow-up procedures are being employed)
  - for incoming queries from respondents (this should not be ignored!)
- computing facilities:
  - hardware, including computers (for word-processing documents, database maintenance, data analysis etc); printers (for printing questionnaires, letters and other documents); scanners (if OMR/OCR questionnaires are being used)
  - software, for word-processing, database management, statistical analysis
  - data entry facilities (commonly, a commercial data preparation agency is used)
- stationery
  - paper on which to print questionnaires and other documents
  - appropriate letterhead (see chapter 6) for covering letters
  - envelopes for outgoing post and for return of completed questionnaires (the latter should be stamped or reply paid, and addressed to the survey organisation; see chapter 6)
  - labels for addressing envelopes (if "window" envelopes are not being used).

Time-tabling a survey

Adequate time must be allowed for each stage in the survey process. The exact time required will depend on whether the questionnaire needs to be developed from scratch, the size of the sample, the mode of administration and the number of pre-notification and follow-up contacts made. There are no hard and fast rules about the timing of dispatch of reminders, but it is sensible to wait until responses to previous mailings have tailed off (generally after 2–3 weeks). It is also useful to specify a cut-off date after which any additional responses will be excluded from the dataset. An example of a timetable for a postal survey is presented in *Table 28*.

It can be seen that a postal questionnaire survey, if carefully designed, prepared and conducted to maximise response and obtain good quality data, can easily take 8 or 9 months from start to finish

and is therefore not a way of obtaining information to answer policy or administrative questions overnight. The need for a survey needs to be anticipated well before the information flowing from it is to be used. The interconnection of tasks – in particular, the dependencies arising from the fact that certain tasks cannot be started until others have been completed – need to be understood, and the tasks on the "critical path" (those tasks that must be started or completed by a certain date if the survey is not to over-run) must be identified. As the survey progresses, actual performance should be compared with the projected schedule, and remedial action taken if necessary.

Costs and costing

As well as time, the resources required to carry out a postal survey will include stationery, postage funding and staff time. Of these, staff time is likely to be by far the most costly item, especially for interviewer-administered surveys, where time for training, travelling between interviews, and field-coding, as well as time spent actually interviewing, needs to be costed. It is important to recognise the "opportunity cost" of taking staff off other tasks, even though, in in-house surveys, these may not be fully recognised in accounting terms. Failure to estimate the staff time requirement realistically is likely to lead to difficulties and contention within the organisation and to skimping of survey tasks, particularly those related to survey quality and quality control.

TABLE 28 *Example of a postal survey timetable*

Weeks	Activity
1–5	Design and pretesting; preparing sample
6–9	Pilot testing
10–11	Refinement and redesign
After 11	Writing data validation and analysis programs
12	Printing
13	Initial mailing
13–14	Processing initial returns
15	1st reminder
17	2nd reminder
15–20	Continue processing returns
21	Dataset closed
21–24	Data coding, checking and cleaning
25–26	Data entry
By 26	Complete writing validation programs
27–28	Data validation and formatting for analysis
By 28	Complete writing analysis programs
29–32	Analysis
33–36	Write up results
37	Finish and celebrate!

It should be noted that the costing of postal surveys is not only a question of identifying operational stages and the types of resource needed at each, it also requires the making of quantitative estimates of unknown factors. For example, estimates must be made of the numbers of sample members who will respond to the initial mailing and to successive reminders, and of the numbers of case records to be processed (particularly at the manual processing stages). Below is shown how stationery and postage requirements for a survey may be estimated. To inform the planning of future surveys, it is good practice to keep careful records of these quantities.

Costing a postal survey: a worked example

In a planned survey, the initial sample size is to be 500. Two reminders are to be used: the first will be a letter; the second will include a duplicate questionnaire as well as a letter. Reply-paid envelopes will be included with the initial mailing and the second reminder. It is assumed that 40% of those contacted will respond to the initial mailing. Of the remaining 60%, 30% will respond to the first reminder, while 40% of the residual 42% will respond to the second reminder. This yields the following numbers of questionnaires to be sent and anticipated to be returned, which gives an overall response rate of 374/500 = 75% (*Table 29*).

Materials needed:

- questionnaires = 500 + 210 (i.e. for initial mailing + 2nd reminder)
- letterhead = 500 + 300 + 210 (i.e. for all mailings)
- large envelopes
  - for posting out = 500 + 210 (i.e. for initial mailing + 2nd reminder)
  - for return post = 500 + 210 (to be enclosed with initial mailing + 2nd reminder)
- small envelopes 300 (i.e. for 1st reminder).

TABLE 29 *Calculating number of questionnaires etc. needed*

Mailing	No. sent	% returned	No. returned
1	500	40	200
2 (1st reminder)	300	30	90
3 (2nd reminder)	210	40	84
Total	1010		374

Costs:

- printing of questionnaires (unit cost × (500 + 210); i.e. for initial mailing + 2nd reminder)
- stationery (envelopes, letterhead; quantities as above)
- postage
  - outgoing = [((500 + 210) × cost of package of questionnaire + letter + envelope) + (300 × cost of letter)]
  - return = (200 + 90 + 84) × cost of questionnaire.

Components of the survey package

The package that the target respondent receives should, as a minimum, comprise a questionnaire and a covering letter (see chapter 6 for more detailed recommendations on the content and style of covering letters). In the case of a postal survey, an envelope in which to return the completed questionnaire to the survey organisation should also be included. In other self-completion questionnaire surveys, provision must also be made for returning questionnaires. In "captive audience" surveys, a field-worker or researcher may gather in the questionnaires at the end of the session (e.g. lesson period for surveys of students). When questionnaires are self-completed in the absence of a field-worker, a box should be provided in which to place them when completed. In both of these situations, providing an envelope in which to place the completed questionnaire prior to handing it in helps to preserve confidentiality.

Postage options for surveys

Although Dillman[17] and others have recommended that stamped rather than franked or reply-paid envelopes should be used for both outgoing and return postage, the review of evidence from primary studies (chapter 6) shows little effect of postage type on response rates, non-response bias or quality of response. Similarly, there is no clear evidence favouring first class postage over second class, although first class obviously offers a slight advantage in terms of speed of delivery. What is most important, however, is that the return postage costs are borne by the survey organisation (either through the use of stamped, addressed envelopes or business reply envelopes) rather than the respondent.

Within the UK, if an item of mail cannot be delivered, for example, because the addressee has moved or the address no longer exists, then a return address, preferably printed on the back of the envelope, will ensure its return to the sender. This is useful in maintaining the accuracy of the mailing list, and there is no charge for returning undelivered items. If the address does not appear on the envelope, Royal Mail personnel will not (in fact, are not allowed to) open the package to ascertain the address of the sender.

There are ranges of facilities offered by the Royal Mail that are of relevance to those conducting postal surveys (details were correct at the time of going to press, but readers are advised to check with the Royal Mail prior to carrying out a survey because the range and cost of facilities are regularly updated).

Mailsort

A contract can be arranged with the Royal Mail to allow discounts on letters posted in bulk, and which have been presorted by postcode. Generally, a minimum of 4000 items in a single mailing is required. Exceptionally, a minimum of 2000 letters will be accepted, provided they are all for delivery within one postcode area. Different levels of service are available:

- first class: target, next day delivery (Mailsort 1)
- second class: target, delivery within 3 days (Mailsort 2)
- economy: target, delivery within 7 days, non-urgent (Mailsort 3).

Cost savings of up to 15% can be earned on Mailsort 1 and 13% on Mailsort 2, while Mailsort 3 savings range between 15% and 25%. These figures refer to 60 g weight items; larger savings apply for heavier items. Savings depend on the service chosen and the level of sorting required. All mailings need to be at least 90% fully postcoded to qualify for entry.

There are also companies who will take an address database and produce lists and labels in the correct order for Mailsort. Some will also stuff envelopes. Costs vary with the number of addresses and the level of service. If the mailing company has a Mailsort contract with the Royal Mail, the survey organisation does not need one.

Printed postage impressions

These are a preprinted alternative to postage stamps or franking machines for outgoing mail to UK destinations. If an organisation mails at least an average 250 inland items a day or spends more than £12,000 a year on postage, it should be eligible to apply to use printed postage impressions (PPIs). However, normal daily post cannot be included in PPI mailing to make up numbers. PPIs may be used only on outgoing mailings to UK

destinations. They must not be used on reply cards, envelopes or labels supplied to respondents. However, there are no weight restrictions and PPIs may be used for recorded deliveries.

Time is saved on preparing the mail because there is no need to frank the envelopes or stick stamps on them. Envelopes or labels can be prepared in advance for special mailings. However, all items in a single PPI mailing must be identical in size, shape and weight. Within a given mailing, all items must also be posted at the same class, which must be indicated on the PPI form; however, the rate of postage can vary from mailing to mailing. PPIs can also be used in conjunction with a wide variety of Royal Mail services, including Mailsort.

Accounting procedures are also simplified by the use of PPIs (e.g. there is no need to "load" a franking machine with prepayment) as an account is opened with the Royal Mail and the user is billed for the postage against the form from the special posting book handed in with each posting. There is a range of options to make payments as convenient as possible, but typically PPIs are invoiced on a monthly basis according to the number of items sent out.

To obtain a PPI licence number, contact the Royal Mail Sales Centre on 0345 950 950 to make an application. Once an application has been approved, a local or headquarters PPI number will be issued, which must then be used on every PPI item. Details of the licence number and office of posting (if appropriate), together with the chosen franking mark, will need to be supplied to a printing company (if used), who will add these to envelopes or labels as requested.

Prepaid envelopes

As an alternative to PPIs, the Royal Mail sell prepaid envelopes with a stamp mark (similar in appearance to a PPI). They are available for both first and second class postage rates. There are three sizes of envelope available. DL envelopes are available in plain or window styles. The types and maximum posting weights for the various sizes are shown below:

- DL (takes A4 folded in three) maximum posting weight 60 g
- C5 (takes A4 folded in two) maximum posting weights 60 g or 100 g
- C4 (takes A4 unfolded) maximum posting weights 60 g or 100 g.

Note that the weight restrictions may mean that these envelopes are generally unsuitable for mailing out questionnaires (although they may be appropriate for letter reminders); in contrast, there are no such weight restrictions on PPIs. For large orders (minimum 5000), an organisational logo can be added to the envelopes.

Business reply service (reply-paid envelopes)

This service enables a person or organisation to receive cards or letters from clients without prepayment of postage. The postage at the first or second class rate, with an additional small handling fee (currently 0.5p) on each item, is paid by the addressee (usually the survey organisation) after return of the items. This means that the organisation pays only for items that are returned, whereas with stamps the costs are incurred regardless of whether the item is sent back or not.

The account for a business reply-paid service must always be in credit. Prior to sending out a survey in which reply-paid envelopes are used, money needs to be transferred into the reply-paid postal account. The Royal Mail will send statements on a regular basis showing the deductions (i.e. what has been spent in terms of the number of items returned and their price) and the remaining credit balance. In addition, the cost of individual items is often written on the outside of the return envelopes by Royal Mail personnel; if it is not, a returned envelope can simply be weighed with the questionnaire inside to calculate the cost. The credit balance generally needs to be kept above £100 by topping up when required. Business reply-paid envelopes can be individualised with the title of the survey, but the statements/invoices from the Royal Mail do not identify specific addressees within a given organisation; so, if multiple surveys are being carried out, an internal accounting system will be required.

A licence to use the service must be obtained from the Royal Mail. A fee is charged to set up this service initially. After obtaining the licence number, the details must be given to whoever will be printing the envelopes and setting up the appropriate plates for printing. Usually, envelopes will need to be supplied.

Business reply can be used for international as well as national mailings. A priority service may be used on payment of a small increased fee to ensure that mail is delivered by the first delivery.

Freepost

A person or an organisation that wishes to obtain a reply from a member of the public without putting

them to the expense of paying postage may, as an alternative to business reply envelopes, include a special Freepost address in the communication. The reply bearing this address can then be posted in the ordinary way, but without a stamp, and the addressee will pay the postage plus a small fee on all the replies that are received. This service is similar to the business reply service but it does not require the use of preprinted envelopes (although these may be used).

Follow-up procedures in surveys

As noted in chapter 6, one of the most powerful means of increasing response rates is through multiple contacts with target respondents. One approach is prenotification: sending a letter or making a telephone call to the sampled individuals in advance of mailing the questionnaire to alert them to its arrival. More usual is the use of reminders. In order to respond, the sampled individual must have both the will and the means to respond. In sending reminders, it is important to address both of these issues.

To stimulate the will to respond, points to consider emphasising (evidence on the most effective approaches is currently lacking, as noted in chapter 6) are as follows:

- Progress to date with the survey: Some respondents will be susceptible to peer pressure and may be encouraged to respond by learning how many people like themselves have already returned a questionnaire.
- The importance of that individual's response: Some people may feel that they personally do not need to bother to respond, especially if response rates to date are good ("What difference will my answers make?"). As with the initial covering letter, it is important to take into account the likely motivation of the study population. It may be worth considering varying the appeal made from that in the initial contact, for instance from an altruistic to a self-interest appeal.
- What to do if the individual has already responded: For example, should they make contact with the survey organisation to check if their questionnaire has been received, or should they simply assume that correspondence has crossed in the post?

Respondents also need the means to respond. Possible reasons for (apparent) non-response is that the original questionnaire was never received, has been damaged or mislaid, or has been lost in the return post. It is therefore a good idea to offer a duplicate questionnaire under these circumstances. However, as identified in chapter 6, research has shown that routinely enclosing a duplicate questionnaire with a first reminder has little effect on response rates to that reminder, but that enclosing it with a second reminder leads to a significantly higher response rate on that occasion.[27]

Researchers should also consider how many reminders to send and at what time intervals. Every reminder will attract more responses (see chapter 6), but at a diminishing rate. Repeated reminders may also cause significant annoyance to those contacted (in health surveys, some research ethics committees believe that more than three contacts constitutes undue pressure: Key L, Newcastle and North Tyneside Joint Research Ethics Committee, Newcastle upon Tyne: personal communication, 1999). Finally, the more reminders are used, the higher the cost and the longer the time required for the survey. Two, or at most three, reminders are probably adequate in most health surveys.

One way of deciding on the timing of reminders is to monitor the rate of return of completed questionnaires and to send the reminders when the rate plateaus. However, this *ad-hoc* approach is at odds with the notion of planning and scheduling, as recommended above. The plateau point will depend to some extent on the length of the questionnaire and on whether first or second class postage is used. Typically, however, the rate of response tends to drop off significantly 2–3 weeks after a given mailing. This pattern suggests that a first reminder should be sent 2–3 weeks after the initial contact, with the second reminder at 4–6 weeks after the original mailing.

Postal survey control systems

The purposes of postal survey control systems are to:

- initiate action (e.g. original mailings and reminders) for each selected individual
- provide up-to-date reports for monitoring overall progress against plans
- allow remedial action to be taken if necessary
- support reports of survey outcome and quality (e.g. response rates; the extent and nature of any non-response bias).

The inputs to the control system are:

- the list of sampled individuals, including allocated identification numbers
- the occurrence and dates of events for each individual (e.g. original mailing, reminders)
- the outcomes of each event (e.g. completed questionnaire returned; notified by the Royal

Mail that a respondent was not known at a specified address).

For very small surveys, a paper-based control system may be used. However, a computerised system, using a database package, is preferable.

The control system should be used as follows.

- During the rehearsal pilot stage: At this point, the system should be set up and tested.
- Before data collection begins: The details of the sampled individuals should be entered and survey numbers assigned.
- When the initial contact is made: The date of initial contact should be recorded.
- On an ongoing daily basis during the data collection period: The outcomes of each event should be recorded; relevant details (e.g. addresses) should be amended as required.
- Periodically (e.g. weekly) during the data collection period: Progress reports (especially response rates to date) should be produced.
- In line with follow-up protocol: Reminders should be generated for non-respondents; the events and dates should be recorded.
- After the close of data collection: The database file should be reconciled with the data files (from entering questionnaires) and a final analysis of response rates and patterns (including, if possible, non-response bias) produced.

Managing dispatch of questionnaires

Managing the dispatch of questionnaires and logging returns need careful attention. For reasons of confidentiality, the names and addresses of target respondents are not usually placed on the questionnaires themselves; rather, a unique identifier (survey number) is used. Consideration needs to be given to how these numbers are added. Sophisticated printing and document reproduction systems allow for the production of "personalised" questionnaires, using a facility akin to the mail-merge function in word-processing packages; however, this is a relatively costly option. A lower-cost alternative is to use a word-processing or database package to produce individualised adhesive labels that can be added to the questionnaires. Similarly, self-inking, automatically advancing number stamps can be used to number questionnaires sequentially. The least satisfactory alternative is to write the survey number on by hand because this approach is error prone. Whichever method is used, care must be taken to place the correct questionnaire and any supporting documentation (e.g. the covering letter) in the right envelope.

Data quality monitoring and data management

Finally, when questionnaires are returned, they need to be carefully checked before data are entered on to a computer for analysis. Things to look out for include: data completeness; inconsistent or implausible answers; and respondents not following instructions (e.g. writing the response in the wrong place or endorsing more than one response when only one is allowed). Any errors should be noted and, if possible, rectified. If open-ended questions have been used, these will need to be coded prior to analysis. Once data have been entered on to the computer, further checks will be required before they are analysed. Points to check for include: if skips have been followed correctly; whether the answers given are valid (e.g. within a sensible range); and if responses are internally consistent (e.g. that a respondent is not claiming to be male and pregnant).

Ethical issues and confidentiality in surveys

Survey researchers need to pay attention to the ethical principles laid down by their profession and the ethical requirements imposed by the survey topic. For example, in health-related surveys, especially those of specific patient groups, approval must be obtained from the local research ethics committee, the members of which will want to see copies of questionnaires, covering letters and other survey documents. It is also important to ensure that all data holdings are registered appropriately in conformance with the Data Protection Act (1998).

Maintaining confidentiality in survey research is most important. As noted in chapter 6, confidentiality means the following:

- Questionnaires are identified only by a survey number; only members of the survey research team can make a link between this survey number and identifiable information (such as names and addresses).
- Access to completed questionnaires is strictly controlled. All those handling questionnaires should be bound by a written code of confidentiality. Completed questionnaires should be stored in locked rooms or cupboards.
- Individual, identifiable responses are not revealed to any third party.
- The results of the survey are not presented in such a way as to allow the identification of any individual respondent.

It is important to note that confidentiality is not synonymous with anonymity.[28] In a completely anonymous questionnaire, no identifier whatsoever (not even a survey number) appears. Anonymity precludes record linkage (e.g. relating a person's questionnaire responses to data on that same individual obtained from another source, such as a medical record). Anonymity also makes sending reminder questionnaires difficult because there is no way of knowing who has responded and who has not. This generally means that reminders have to be sent to everyone, which can cause annoyance and confusion to those who have already responded, and can result in the receipt of two completed questionnaires (often with different responses!) from some participants. One way around this, if anonymity is believed to be essential, is to ask respondents to send back an identifiable letter or postcard under separate cover,[29] so that the survey researchers can keep track of who has and who has not responded (although they will be unable to work out which questionnaire came from which respondent).

Further sources of information on good practice in survey research

Finally, there are a number of key text books on survey research that complement and supplement this review (e.g.[4,17,24,26,30–35]).

Useful websites for survey researchers include:

http://www.natcen.ac.uk/
http://sites.netscape.net/gsociology/methods/
http://www.princeton.edu/~abelson/xpractic.html
http://www.ukans.edu/cwis/units/coms2/po/index.html

Finally, methodological reports from well-designed health-related surveys (e.g.[36,37]) provide examples of good practice in surveys of patients and the general population.

References for appendix 1

1. Kish L. Survey sampling. New York: Wiley; 1965.
2. Cochran WG. Sampling techniques. New York: Wiley; 1977.
3. Barnett V. Elements of sampling theory. London: Hodder and Stoughton; 1974.
4. Fowler FJ. Survey research methods. 2nd ed. Beverly Hills, CA: Sage; 1993.
5. McColl E, Thomas R. The use and design of questionnaires. London: Royal College of General Practitioners; 2000.
6. Touw-Otten F, Meadows K. Cross-cultural issues in outcome measurement. In: Hutchinson A, McColl E, Christie M, Riccalton C, editors. Health outcome measures in primary and out-patient care. Amsterdam: Harwood Academic; 1996. p. 199–208.
7. Bentzen N, Christiansen T, McColl E, Meadows K. Selection and cross-cultural adaptation of health outcome measures. *European Journal of General Practice* 1998;**4**:27–33.
8. McDowell I, Newall C. Measuring health: a guide to rating scales and questionnaires. Oxford: Oxford University Press; 1987.
9. Wilkin D, Hallam L, Doggett M-A. Measures of need and outcome for primary health care. Oxford: Oxford University Press; 1992.
10. Bowling A. Measuring disease. Buckingham: Open University Press; 1995.
11. Bowling A. Measuring health: a review of quality of life measurement scales. 2nd ed. Buckingham: Open University Press; 1997.
12. Fitzpatrick R, Davey C, Buxton MJ, Jones DR. Evaluating patient-based outcome measures for use in clinical trials. *Health Technology Assessment* 1998;**2**(14).
13. Campanelli PC, Martin EA, Rothgeb JM. The use of respondent and interviewer debriefing studies as a way to study response errors in survey data. *The Statistician* 1991;**40**:253–64.
14. Oksenberg L, Cannell CF, Kalton G. New strategies for pretesting survey questions. *Journal of Official Statistics* 1991;**7**:349-65.
15. Schwarz N, Bradburn NM, Sudman S. Thinking about answers: the application of cognitive processes to survey methodology. San Francisco, CA: Jossey-Bass; 1996.
16. Schwarz N, Sudman S. Answering questions: methodology for determining cognitive and communicative processes in survey research. San Francisco, CA: Jossey-Bass; 1996.
17. Dillman DA. Mail and telephone surveys: the total design method. New York: Wiley; 1978.
18. Brown TL, Decker DJ, Connelly NA. Response to mail surveys on resource-based recreation topics: a behavioral model and an empirical analysis. *Leisure Sciences* 1989;**11**:99–110.
19. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. Oxford: Oxford University Press; 1989.
20. Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin* 1959;**56**:81–105.
21. Cronbach L. Coefficient alpha and the internal structure of tests. *Psychometrika* 1951;**16**:287–334.

22. Nunnally J, Bernstein JC. Psychometric theory. 3rd ed. New York: McGraw-Hill; 1994.

23. Morton Williams J. Interviewer approaches. Aldershot: Dartmouth; 1993.

24. Salant P, Dillman DA. How to conduct your own survey. New York: Wiley; 1994.

25. Frey JH, Oishi SM. How to conduct interviews by telephone and in person. Thousand Oaks, CA: Sage; 1995.

26. Fink A. The survey kit. Thousand Oaks, CA: Sage; 1995.

27. Swan JE, Epley DE, Burns WL. Can follow-up response rates to a mail survey be increased by including another copy of the questionnaire? *Psychological Reports* 1980;**47**:103–6.

28. Zeinio RN. Data collection techniques: mail questionnaires. *American Journal of Hospital Pharmacy* 1980;**37**:1113–19.

29. Sudman S. Mail surveys of reluctant professionals. *Evaluation Review* 1985;**9**:349–60.

30. Moser CA, Kalton G. Survey methods in social investigation. 2nd ed. Aldershot: Gower; 1971.

31. de Vaus DA. Surveys in social research. 3rd ed. London: UCL Press; 1991.

32. Oppenheim AN. Questionnaire design, interviewing and attitude measurement. 2nd ed. London: Pinter; 1992.

33. Mangione TW. Mail surveys – improving the quality. Thousand Oaks, CA: Sage; 1995.

34. Fink A, Kosecoff J. How to conduct surveys: a step-by-step guide. Thousand Oaks, CA: Sage; 1998.

35. Dillman DA. Mail and internet surveys: the tailored design method. 2nd ed. New York: Wiley; 2000.

36. National Centre for Social Research, Picker Europe, Department of Primary Health Care and General Practice. National surveys of NHS patients: General practice 1998. London: NHS Executive; 1999.

37. Erens B, Primatesta P, Prior G. Health survey for England: the health of minority ethnic groups 1999. London: National Statistics; 1999.

Appendix 2

Topics for data abstraction

CODE	TOPIC
1. Method of administration	
1.1	*Face-to-face interviews (nos)*
1.1.1	'Normal' face-to-face interviews
1.1.2	Computer-assisted face-to-face interviews
1.2	*Self-completion questionnaires (nos)*
1.2.1	Postal questionnaires/paper-based self-completion questionnaires
1.2.2	Other means of distribution (e.g. captive audience)
1.2.3	Computer-assisted self-completion questionnaires
1.3	*Telephone interviews (nos)*
1.3.1	'Normal' telephone interviews
1.3.2	Computer-assisted telephone interviews
2. Sources of bias of particular relevance in interview surveys	
2.1	*Interviewer effects (nos)*
2.1.1	Gender of interviewer
2.1.2	Age of interviewer
2.1.3	Social class of interviewer
2.1.4	Race/ethnicity of interviewer
2.1.5	Experience/training of interviewer
2.1.6	Other characteristics of interviewer
2.2	*Remuneration of interviewers*
2.2.1	Hourly rate vs. rate per completed interview
2.3	*"Environment/context" effects*
2.3.1	Prior notification of interview vs. cold calling
2.3.2	Place of interview (e.g. hospital vs. home)
2.3.3	Presence of others (e.g. spouse, significant other)
2.3.4	Prior experience of being interviewed
3.	**Issues to do with questionnaire appearance**
3.1	*Booklet vs. tatty "stapled through top left corner"!*
3.2	*Size of page (e.g. A4 vs. smaller)*
3.3	*Double-sided vs. single-sided printing*
3.4	*Size of print*
3.5	*Length*
3.5.1	Number of questions
3.5.2	Number of pages
3.6	*Colour of paper*
3.7	*Impact of illustrations/visual aids*
3.8	*"Tick boxes" vs. "circle the number" response format*
3.9	*Codes before or after descriptors*
3.10	*Instructions on questionnaire vs. in covering letter*

4. Methods of trying to enhance response rates for questionnaire

4.1	*Disclosure of sponsorship*
4.2	*Anonymity vs. confidentiality*
4.3	*Distribution strategies*
4.3.1	Personal delivery vs. postal
4.3.1.1	Colour of envelope
4.3.1.2	1st or 2nd class mail
4.3.1.3	Stamped vs. franked envelope
4.3.1.4	Denominations/types of stamps
4.3.2	Covering letter
4.3.2.1	No letter vs. letter
4.3.2.2	Personalised vs. generalised salutation
4.3.2.3	Personalised vs. generalised (e.g. rubber stamp) signature
4.3.2.4	Characteristics of person signing letter (e.g. gender, status)
4.4	*Return mechanisms*
4.4.1	To "sponsor" or to "neutral location"
4.4.2	Personal collection vs. "drop box" vs. postal
4.4.3	Reply paid vs. stamped envelopes
4.4.3.1	1st vs. 2nd class mail
4.4.3.2	Stamped vs. reply-paid envelope
4.4.3.2	Denomination/type of stamp
4.5	*Incentives (financial and other)*
4.5.1	None/"something for all"/"prize draw"
4.6	*Reminders*
4.6.1	Number of reminders
4.6.2	Timing of reminders
4.6.3	Use of duplicate questionnaire
4.6.4	Mode of contact (postcard/letter/phone/personal)
4.7	*Time of year of distribution*
4.8	*Specification of deadline dates*

5. Question wording/sequencing/response categories

5.1	*Open-ended vs. closed questions*
	For open-ended questions ...
5.1.1	Lines or blank space to record answers
5.1.2	Amount of space allowed
	For closed questions ...
5.1.3	Ordering of response options
5.1.4	Number of response categories
5.1.4.1	How many points in attitude scales (e.g. 5 vs. 7)
5.1.4.2	Odd vs. even number of points
5.1.4.3	Neutral mid-point in attitude scales
5.1.4.4	Does attaching numbers alter response?
5.1.4.5	Does ordering of numbers alter response?
5.1.4.6	Offering "other", "don't know", "not applicable" as explicit options
5.1.5	Should verbal labels be used or just numbers?
5.2	*Question sequencing*
5.2.1	Influence of filtering
5.2.2	Influence of context (e.g. placement of general vs. specific questions)

5.3	*Frames of reference for question wording*
5.3.1	Direct vs. indirect
5.3.2	Time-frames (open vs. specific)
5.3.3	Personal vs. general
5.3.4	Implicit vs. explicit alternatives
5.4	*Language*
5.4.1	Specialist vs. general
5.4.2	Vocabulary and reading age
5.4.3	Negative vs. positive wording
5.5	*For longitudinal studies, whether previous round responses should be fed back*

6. Handling 'sensitive' questions

6.1	*"Casual" approach*
6.2	*Numbered card*
6.3	*"Everybody" approach*
6.4	*"Other people" approach*
6.5	*Sealed ballot*
6.6	*"Kinsey" technique*
6.7	*Two-question technique*

Appendix 3

Data abstraction form

Designing and using patient and staff questionnaires: a review of best practice
Data abstraction form

0. Does the content of the paper meet the study inclusion criteria?

Yes → Complete this data extraction form
No → Do not complete this form; log on "excluded studies" form

1. Report identification and key characteristics

i Ref. no.:

ii Title:

iii Reviewer: (please circle) CB AJ EMcC LT JS NS

iv Focus of study: (please tick)

Health ☐
Non-health ☐

v Study design: (please tick)

1 Randomised controlled trial ☐
2a Non-random concurrent controlled study ☐
2b Self-controlled study ☐
2c Historically controlled study ☐
3a Cross-sectional study ☐
3b Cohort study ☐
3c Case-control study ☐
4a Meta-analysis, with systematic review ☐
4b Systematic review without meta-analysis ☐
4c Meta-analysis, with non-systematic review ☐
4d Non-systematic review without meta-analysis ☐
5a Theoretical paper ☐
5b Position paper ☐

2. Key words for topics covered by this paper (please circle all that apply)

1.	2.	3.	4.	5.	6.
1.1	*2.1*	*3.1*	*4.1*	*5.1*	*6.1*
1.1.1	2.1.1			5.1.1	
1.1.2	2.1.2			5.1.2	
	2.1.3			5.1.3	
	2.1.4			5.1.4	
				5.1.4.1	
				5.1.4.2	

continued

contd

1.	2.	3.	4.	5.	6.
				5.1.4.3	
				5.1.4.4	
				5.1.4.5	
				5.1.4.6	
	2.1.5			5.1.5	
	2.1.6				
1.2	*2.2*	*3.2*	*4.2*	*5.2*	*6.2*
1.2.1	2.2.1			5.2.1	
1.2.2				5.2.2	
1.2.3					
1.3	*2.3*	*3.3*	*4.3*	*5.3*	*6.3*
1.3.1	2.3.1		4.3.1	5.3.1	
			4.3.1.1		
			4.3.1.2		
			4.3.1.3		
			4.3.1.4		
1.3.2	2.3.2		4.3.2	5.3.2	
			4.3.2.1		
			4.3.2.2		
			4.3.2.3		
			4.3.2.4		
	2.3.3			5.3.3	
	2.3.4			5.3.4	
		3.4	*4.4*	*5.4*	*6.4*
			4.4.1	5.4.1	
			4.4.2	5.4.2	
			4.4.3	5.4.3	
			4.4.3.1		
			4.4.3.2		
		3.5	*4.5*	*5.5*	*6.5*
		3.5.1	4.5.1		
		3.5.2			
		3.6	*4.6*		*6.6*
			4.6.1		
			4.6.2		
			4.6.3		
			4.6.4		
		3.7	*4.7*		*6.7*
		3.8	*4.8*		
		3.9			
		3.10			

DECISION POINT 1

Does the author(s) make any recommendations for future research?

Yes → Complete Section 3
No → Proceed directly to Decision Point 2

3. Recommendations for future research

DECISION POINT 2

Is this a descriptive study, review article or theoretical paper (coded as study types 3a–5b at 1.v above)?

Yes → No further data abstraction is needed at this time. Remember to highlight relevant post-1975 references (Section 7).
No → Continue with Section 4.

4. Methodological criteria for inclusion as evidence (please circle one response on each line)

i	Minimum group size ≥50 at outset	Yes	No
ii	Subjects randomly allocated to groups	Yes	No
iii	Control and intervention groups comparable at baseline	Yes	No
iv	Methodological intervention stated	Yes	No
v	Methodological intervention evaluated	Yes	No

DECISION POINT 3

Have you answered **Yes** for items i, iv and v above, and answered Yes for either item ii or item iii?

Yes → Continue with Section 5.
No → No further data abstraction is needed at this time. Remember to highlight relevant post-1975 references (Section 7).

5. Rating of study quality (please circle one response on each line)

Study population				
Inclusion criteria stated?	Yes		No	
Exclusion criteria stated?	Yes	?	No	
Respondents randomly allocated to control and intervention groups?	Yes	?	No	
Comparability of control and intervention groups at baseline examined?	Yes	?	No	
Control and intervention groups comparable at baseline?	Yes	?	No	Not applicable
Methodological intervention				
Underlying theoretical orientation or empirical justification stated?	Yes	?	No	
Control and intervention groups treated equally apart from intervention?	Yes	?	No	
Measurement of effects				
Primary outcome measure(s) described?	Yes	?	No	
Outcomes measured in common units?	Yes	?	No	
Instrument response rates for control and intervention groups reported?	Yes	?	No	
Item response rates for control and intervention groups reported?	Yes	?	No	

continued

contd

Non-response bias in control and intervention groups reported?	Yes	?	No
Financial costs of control and intervention methodologies reported?	Yes	?	No
Analysis and reporting			
All respondents accounted for?	Yes	?	No
Method of analysis stated?	Yes	?	No
Information on statistical significance of effect?	Yes	?	No
Information presented on size of effect?	Yes	?	No
Information presented on precision of effect?	Yes	?	No

6. Summary of study

DESIGN	
Methodological interventions:	
Setting:	
Country:	
Study population:	Group sizes: Inclusion criteria: Exclusion criteria:

Primary outcome measures:	
Instrument response rates	☐
Instrument completion rates	☐
Item response rates	☐
Non-response bias	☐
Scores on specified scales	☐
Financial costs	☐
Other (please specify)	☐

RESULTS	
Instrument response rates:	Number of subjects recruited: Response rates: Significance of difference in response rates:

continued

contd

Instrument completion rates:	Number of participants recruited: Response rates: Significance of difference in response rates:
Item response rates:	Number of items in questionnaire: Response rates: Significance of difference in response rates:
Non-response bias:	With respect to ... Gender: Age: Socio-economic status: Education: Other factors:
Scores on specified scales:	Scale scores: Significance of difference in scale rates:
Financial costs:	Costs for intervention and control groups: Significance of difference in costs:
Other primary outcome measures:	Summary of results: Significance of difference in results:
Conclusions:	
Recommendations for practice:	

7. Administrative details

Relevant references highlighted by reviewer	Yes	No
Highlighted references checked on REFMAN	Yes	No
Extra references identified and sought	Yes	No
Details from data extraction form entered into Access	Yes	No

Appendix 4

Data abstraction manual

Designing and using patient and staff questionnaires: a review of best practice

Data abstraction codebook

0. Does the content of the paper meet the study inclusion criteria?
(see section on "Inclusion and exclusion criteria" below)

If "**Yes**": Complete data abstraction form
If "**No**": Do not complete data abstraction form
Record ref. ID and reason for exclusion on "excluded studies" form
Keep reference on REFMAN

1. Report identification and key characteristics

i	Ref. no.	This is the unique identifier allocated to the reference by REFMAN (i.e. its ref. ID).
ii	Title	Copy the title of the paper as a double check on the REFMAN ID.
iii	Reviewer	Initials of reviewer: CB, AJ, EMcC, LT, JS, NS
iv	Focus of study	Define as health or non-health; code surveys of healthcare professionals and student members of these professions as health
v	Study design	Code as: 1 Randomised controlled trial 2a Non-random concurrent controlled study 2b Self-controlled study 2c Historically controlled study 3a Cross-sectional study 3b Cohort study 3c Case-control study 4a Meta-analysis, based on systematic review 4b Systematic review without meta-analysis 4c Meta-analysis, based on non-systematic review 4d Non-systematic review without meta-analysis 5a Theoretical paper 5b Position paper (for explanation see section on "Study designs" below))

2. Key words for topics covered by this paper

Circle code numbers for all topics covered by this paper (not just the primary focus). Code at the most specific level possible [code numbers as in appendix 2 of this report]. Thus, a study of whether to include a duplicate questionnaire with reminders should be coded as "duplicate questionnaire" (code number 4.6.4) rather than simply as "reminders" (code number 4.6). Code as much detail as possible. For example, all studies involving a postal self-completion questionnaire should be coded 1.2.1, while all those involving standard face-to-face interviews should be coded as 1.1.1. (If you feel that additional codes are required, please see Elaine McColl.)

DECISION POINT 1

Does the author make any recommendations for future research?

If "**Yes**", complete Section 3.
If "**No**", proceed directly to Decision Point 2.

3. Recommendations for future research

Include **only** recommendations for future research made by the **author** of this paper. Insofar as is possible, record these recommendations verbatim, in quotation marks, citing the relevant page reference. Do not record ideas for future research that are apparent to you as a reviewer but are not mentioned explicitly in the paper.

DECISION POINT 2

Is this a descriptive study, review article or theoretical paper (coded as study types 3a–5b at 1.v above)?

If "**Yes**", no further data abstraction is needed at this time. However, check to ensure that all relevant references from 1975 on have been highlighted (see Section 7).
If "**No**", continue with Section 4.

4. Methodological criteria for inclusion of study as evidence

This section should be completed for all comparative studies.

i	Minimum group size ≥50 at time of allocation	To be rated as "Yes", each intervention group and control group must contain at least 50 participants at the point of randomisation or other means of allocation to groups. For factorial designs, there should be at least 50 participants per comparison group for each of the main effects at the point of allocation to groups. For all designs, rate as "No" if group sizes were less than 50 at the point of group allocation.
ii	Participants randomly allocated to groups	If the paper contains a statement that participants were allocated to the control and intervention groups randomly, rate as "Yes". Also rate as "Yes" if the entire study sample was selected randomly, although individuals may have been allocated to control and intervention groups in a systematic fashion. If it is unclear whether random allocation was used, or the method of allocation is not stated, or the method was non-random (e.g. systematic sampling from an alphabetical list) rate as "No".
iii	Control and intervention groups comparable at baseline	If the paper contains information (either in the text or in tabular form) showing that the control and intervention groups were comparable in terms of relevant characteristics at baseline, rate as "Yes". If no comparison of groups at baseline was carried out, or if it is not clear whether the groups were comparable, rate as "No".
iv	Methodological intervention stated	If the specific aspect(s) of survey methodology that is/are the focus of the intervention is/are stated, rate as "Yes". Otherwise rate as "No".
v	Methodological intervention evaluated	If the paper contains an evaluation of the impact of the methodological intervention, rate as "Yes". Otherwise rate as "No".

DECISION POINT 3

Is the study rated as "**Yes**" on items **i**, **iv** and **v** above, and "**Yes**" on **either** item **ii** or item **iii**?

If "**Yes**", continue with section 5.
If "**No**", no further data abstraction is needed at this time. However, check to ensure that all relevant references from 1975 on have been highlighted (see Section 7).

5. Rating of study quality

This section should be completed for all comparative studies fulfilling the methodological inclusion criteria (Section 4).

Study population	
Inclusion criteria stated?	If the author provides any details of the population eligible for the study, code as "Yes".
	If no details of the eligible population are given, code as "No".
Exclusion criteria stated?	If the author specifies any subgroups of the population who were not eligible for the study, code as "Yes". Similarly, if it is stated that there were no exclusions, code as "Yes".
	If exclusions are mentioned but details of their characteristics are not given, code as "?".
	If no mention of exclusions are made, code as "No".
Participants randomly allocated to control and intervention groups?	If the method of randomly allocating participants to groups (e.g. random numbers tables; sealed envelopes etc.) is specified, code as "Yes".
	If the author simply states "random allocation" without specifying the method, code as "?".
	If allocation was non-random (e.g. systematic), code as "No".
Comparability of control and intervention groups at baseline examined?	If the author explicitly states that a baseline comparison of relevant characteristics of the control and intervention group was carried out, code as "Yes". Similarly, if results are presented to indicate implicitly that such a comparison was performed, code as "Yes".
	If it is not explicitly stated that no comparison was carried out but there is no evidence that the groups were compared at baseline, code as "?".
	If there is an explicit statement that no comparison of groups at baseline was carried out, code as "No".
Control and intervention groups comparable at baseline?	If it is clear that no comparison of baseline characteristics was performed (i.e. previous question has been coded as "No"), code as "not applicable".
	If the author presents information (either in the text or in tabular form), showing that the control and intervention groups were comparable in terms of relevant characteristics at baseline, code as "Yes".
	If a comparison was carried out, but the results are not reported, code as "?". Similarly, if there is no evidence of a comparison being performed (i.e. previous question has been coded as "?"), code as "?".
	If it is stated that analysis showed that the groups were not comparable at baseline, code as "No".

continued

contd

Methodological intervention	
Underlying theoretical orientation or empirical justification stated?	If a theoretical explanation as to why the intervention might be expected to influence outcome is presented or if the intervention is based on previous empirical work, code as "Yes". If it is unclear whether or not the intervention is based on an underlying theoretical position or on previous empirical evidence, code as "?". If it is explicitly stated that the intervention was developed arbitrarily and that there was no theoretical reason or empirical basis for suspecting that it might influence outcome, code as "No".
Control and intervention groups treated equally apart from intervention?	If the author states that other confounding interventions were avoided and that, apart from the methodological intervention, control and intervention groups were treated equally, code as "Yes". If it is not clear whether treatment of groups was equal, code as "?". If the author specifies other confounding interventions or differences in treatment of groups, code as "No".
Measurement of effects	
Primary outcome measure(s) described?	If the author clearly specifies the primary outcome measure(s) used and gives sufficient detail about the measure(s) to permit replication, code as "Yes". If some information is given about the primary outcome measure(s), but not sufficient to permit replication, code as "?". If no details of the primary outcome measure(s) are given, code as "No".
Outcomes measured in common units?	If the author states outcomes were measured in identical units for control and intervention groups (e.g. final response rates, mean scores on the same scale), code as "Yes". Similarly, if the outcomes were adjusted to common units (e.g. if cost data were collected in 1993 for the control group and in 1995 for the intervention group, but costs have been adjusted to 1995 levels for the control group), code as "Yes". If outcomes were measured in different units, but sufficient information is presented to permit adjustment to common units (e.g. if it is clear that costs for the control group are at 1993 rates and for the intervention group at 1995 rates, thereby allowing adjustment using an appropriate inflation factor), code as "?". If outcomes were not measured in identical units (e.g. response rates at 6 weeks for the intervention group and at 4 weeks for the control group), and have not been adjusted, and there is insufficient information presented to allow adjustment, code as "No".
Instrument response rates for control and intervention groups reported?	Instrument response rates refer to the total % of questionnaires returned, which may include those returned blank. If separate instrument response rates are presented for control and intervention groups (either in the text or in tables), code as "Yes".

continued

contd

	If only an overall (i.e. for control and intervention groups combined) instrument response rate is reported, but it is stated that there was no significant difference in instrument response rates between control and intervention groups, code as "?".
	If only an overall instrument response rate is reported, and there is no information at all on separate instrument response rates for control and intervention groups, code as "No".
Item response rates (or item omission rates) for control and intervention groups reported?	These data may be presented in terms of the complement of item response rates, in other words, item omission rates. This is acceptable; the questions below may be answered with respect to item omission rates.
	If separate item response rates (for one or more questions in the instrument) are presented for control and intervention groups (either in the text or in tables), code as "Yes".
	If only an overall (i.e. for control and intervention groups combined) item response rate is reported, but it is stated that there was no significant difference in item response rates between control and intervention groups, code as "?".
	If only an overall item response rate is reported, and there is no information at all on separate item response rates for control and intervention groups, code as "No".
Non-response bias in control and intervention groups reported?	Non-response bias in this context refers to the difference between respondents and non-respondents. Authors may have used a different interpretation of this term, but this question should be answered with respect to our definition.
	If information on non-response bias is presented separately for control and intervention groups (either in the text or in tables), code as "Yes".
	If only information on overall non-response bias is reported but it is stated that there was no significant difference in non-response bias between control and intervention groups, code as "?".
	If only information on overall non-response bias is reported, and there is no information at all on non-response bias in control and intervention groups separately, code as "No".
Financial costs of control and intervention methodologies reported?	If separate financial costs are presented for control and methodologies (either in the text or in tables), code as "Yes".
	If only overall financial costs are reported but it is stated that there was no significant difference in financial costs between control and intervention methodologies, code as "?".
	If only overall financial costs are reported, and there is no information at all on financial costs of control and intervention methodologies separately, code as "No".
Analysis and reporting	
Are all participants accounted for?	If all participants who were recruited to the study are accounted for in the results, code as "Yes".
	In a longitudinal study, if the number surveyed and/or responding at each time point is not specified, code as "?".

continued

contd

	If there is an unexplained discrepancy between the number recruited and the number for whom data are presented, code as "No".
Method of analysis stated?	If the author explicitly states all statistical tests used, code as "Yes". (Note that simply reporting an *F* or *t* statistic in the results is not indicative of the statistical method used.)
	If details of some but not all statistical tests used are given, code as "?".
	If no details of statistical tests used are given, code as "No".
	(Note that no judgement is to be made as to the appropriateness of the tests chosen.)
Information presented on statistical significance of effect?	If the statistical significance (i.e. *p*-value) of observed differences in one or more of the primary outcome measure(s) with respect to at least one of the factors under investigation is given, code as "Yes".
	If the statistical significance of some, but not all, outcome measures is given, code as "?".
	If no information is given on the statistical significance of any outcome measures, code as "No".
Information presented on size of effect?	If the value (e.g. mean score on a scale, % response rate) of the primary outcome measure(s) is presented separately for control and intervention groups, so that the reader can calculate the size of the effect, code as "Yes". *Example:* "Response rates were 80% for those interviewed and 75% for those receiving a postal questionnaire."
	If information on the size of effect is given (e.g. mean differences on scale score, % difference in response rate), but information on the value of the primary outcome measure(s) is not given separately for control and intervention groups, code as "?". *Example:* "The response rate for those interviewed was 5% higher than for those receiving a postal questionnaire."
	If no information is provided on the size of the effect for the primary outcome measure(s) for control and intervention groups, code as "No".
Information presented on precision of effect?	If there is a clearly stated interval estimate of the effect size (e.g. a specific confidence interval or standard error), code as "Yes". *Examples:* "The effect size was +7% with a 95% CI, 3 to 21." or "The effect size was X with standard error Y".
	If an ambiguous interval estimate of the precision of the effect is presented, code as "?". *Examples:* "The effect size was between 3% and 21%" (with no indication of whether these are 95% or 99% CIs). or "The effect was X ± Z" (where it is unclear whether Z is one standard error or some multiple of it).
	If no interval estimate of effect size is given, code as "No".

6. Summary of study

This section should be completed for all comparative studies fulfilling the methodological inclusion criteria as evidence (Section 4).

Summary of study	
Methodological interventions	The main comparisons (effects) examined in this study (e.g. face-to face interview vs. questionnaire; personalised vs. generalised salutation).
Setting	Specify as: None stated Hospital – nos Hospital – inpatient Hospital – outpatient Community – nos Community – primary care Community – residential home Community – own home Community – work place Education establishment – nos Educational establishment – 3rd level Educational establishment – 2nd level Educational establishment – 1st level Professional organisation Other (please specify)
Country	Specify as: None stated UK Other European USA Canada Australia New Zealand Other (please specify)
Study population	Describe the study population, including the size of each group at point of entry to the study, and any inclusion/exclusion criteria.
Primary outcome measure(s)	Specify the measure(s) used. Specify as: Instrument response rates Instrument completion rates Item response rates/item omission rates Non-response bias (respondents vs. non-respondents) Scores on specified scales (e.g. health status, quality of life) Financial costs Other (please specify)
Results	
Instrument response rates	Instrument response rates refer to the total % of questionnaires returned, which may include those returned blank. This section should be completed only if instrument response rates were examined in the paper.

continued

contd

	If the information is contained in the paper, report the number of participants recruited to control and intervention groups, together with response rates in % terms (for control and intervention groups separately, if possible), and indicate whether there was a significant difference in response rates between control and intervention groups, stating *p*-values where quoted.
Instrument completion rates	Instrument completion rates refer to the % of questionnaires returned in a completed state. It excludes those returned blank and authors may define it to exclude certain other categories (e.g. those with more than X% missing data; those returned after a cut-off date). This section should be completed only if instrument completion rates were examined in the paper. If the information is contained in the paper, include the number of participants recruited to control and intervention groups, together with completion rates in % terms (for control and intervention groups separately, if reported), and indicate whether there was a significant difference in completion rates between control and intervention groups, stating *p*-values where quoted.
Item response rates	These outcomes may be presented in terms of the complement of item response rates, in other words, item omission rates. This is acceptable; in such cases report findings in terms of item omission rates, specifying that this is what is being considered. This section should be completed only if item response rates were examined in the paper. If the information is included in the paper, report the total number of items, a measure of central tendency (e.g. mean, median) of item response rates, and a measure of the variability in item response rates (e.g. range, standard deviation). If possible, report these values separately for control and intervention groups, and indicate whether there is a significant difference in item response rates between control and intervention groups, stating *p*-values where quoted.
Non-response bias	Non-response bias in this context refers to the difference between respondents and non-respondents. Authors may have used a different interpretation of this term. This section should be completed only if non-response bias according to the definition above (e.g. differential response rates with respect to demographic characteristics) was examined in the paper. If the information is given in the paper, include details of non-response bias within control and intervention groups and indicate whether there is a significant difference in non-response bias between control and intervention groups. Where available, give results separately for: gender, age, socio-economic status (socio economic group, social class) and level of education and any other variables considered. Specify *p*-values where quoted.
Scores on specified scales	This section should be completed only if scores on specified scales (e.g. on health status/quality of life scales) were examined/reported in the paper.

continued

contd

	If the information is included in the paper, report a measure of central tendency (e.g. mean, median) of scale scores and a measure of the variability in these scores (e.g. range, standard deviation). If possible, report these values separately for control and intervention groups, and indicate whether there is a significant difference in item response rates between control and intervention groups, quoting *p*-values where available.
Financial costs	This section should be completed only if financial costs of the intervention were examined/reported in the paper. If the information is included in the paper, report a measure of central tendency (e.g. mean, median cost) of financial costs and a measure of the variability in these costs (e.g. range, standard deviation). If possible, report these values separately for control and intervention groups, and indicate whether there is a significant difference in item response rates between control and intervention groups, quoting *p*-values where available.
Other primary outcome measures	This section should be completed only if the author examined any primary outcomes not listed above. If the information is included in the paper, report a measure of central tendency (e.g. mean, median cost) of these outcomes and a measure of the variability in these outcomes (e.g. range, standard deviation). If possible, report these values separately for control and intervention groups, and indicate whether there is a significant difference in item response rates between control and intervention groups, quoting *p*-values where available.
Conclusions	List the main conclusions as stated in the paper.
Recommendations	List the main recommendations for practice as stated in the paper.

7. Administration

Any extra potentially relevant references identified?	Reviewer should highlight any references in the list at the end of the paper (published books and papers, but not "grey" literature such as internal reports of research organisations or unpublished theses) that may be potentially relevant to this review. Only references published from **1975 onwards** should be selected.
Highlighted references checked on REFMAN	In order to avoid obtaining duplicate references, REFMAN should be searched (by project secretary) for each highlighted reference to check whether we already have it. If not, consult with Elaine McColl before seeking the reference.
Extra references identified and sought	Extra references should be identified and obtained as soon as possible (by project secretary).

Inclusion and exclusion criteria

Inclusion criteria

Types of paper/levels of evidence

Levels of evidence will be graded from the highest level of evidence from experimental designs (in particular, randomised controlled trials) to lower-level evidence from single-approach studies. Evidence will be gathered from both the health sector and other sectors (e.g. education research, market research) because methodological messages are likely to be generalisable across sectors. The next section on "Study designs" indicates the overall hierarchy of evidence (but we recognise that a high-quality non-random design may provide better evidence than a poor-quality randomised controlled trial). The following types of paper will be considered:

- papers reporting on experimental designs (in particular, randomised controlled trials) in which two or more approaches to questionnaire design and/or administration are compared
- papers reporting on other comparative studies in which two or more approaches to questionnaire design and/or administration are compared
- papers reporting on theoretical or empirical studies in which the advantages and disadvantages of a single approach to questionnaire design and/or administration are reported
- review articles and position papers; these will be accessed mainly as a potential source of useful references and of suggestions for further research.

Focus of papers

- The key question is "does this work help to identify best practice with respect to survey design and administration"? If the answer is "Yes", the paper should be included.
- Reference should be made to the keyworded topic list in deciding on whether to include a specific paper.
- Studies of particular types of questions (e.g. situational response, circular questions) should be included, particularly if these are novel types of questions.
- Studies of particular approaches to structured interviewing (in particular, cognitive interviews) should be included if these approaches are being used in the context of research (but not if they are being used in other contexts, e.g. police interviews).
- Comparisons of short form vs. long form measures should be included **only** if the focus of the paper is the effect on response rates or other key issues identified in the topic list. They should not be included if the focus is solely on the reliability/validity of the information yielded (see "Exclusion criteria" below).
- Comparisons of proxies vs. self-report should be included **only** if the focus of the paper is the effect on response rates or other key issues identified in the topic list. They should not be included if the focus is solely on the reliability/validity of the information yielded (see "Exclusion criteria" below).
- The effect of revealing researchers' bias should be included.
- The impact of interviewer characteristics (e.g. age, gender) on response rates, response bias etc. should be included.
- Comparisons of questionnaires/interviews vs. diary methods should be included.
- Comparisons of rating scales vs. questionnaire scales should be included.
- Techniques for handling sensitive topics should be included.
- Papers on individual tailoring of questionnaires to respondents should be included.
- Papers on the context (e.g. patient groups, subject matter) in which interactive computerised techniques are employed (even if these are only single-approach studies) should be included; we believe this approach to be relatively novel and therefore unlikely to have received much coverage in the classic texts.
- Papers on the context in which video questionnaires are employed (even if these are only single-approach studies) should be included; we believe this approach to be relatively novel and therefore unlikely to have received much coverage in the classic texts.

Exclusion criteria

- Studies of qualitative approaches to data collection (e.g. unstructured interviews, focus group discussions) should be excluded; qualitative methods form the focus of another review funded under the HTA Methodology initiative.
- Comparisons of structured (questionnaire) approaches with unstructured (qualitative) approaches should be excluded.
- Studies of the use of questionnaires specifically in the context of Delphi surveys or other consensus methods should be excluded; consensus methods form the focus of another review funded under the HTA Methodology initiative.
- Studies of the applicability of the questionnaire approach (i.e. whether questionnaires are a feasible method of data collection) to particular topics (e.g. assessment of diet or smoking behaviour) should be excluded. Although we recognise the initial choice of questionnaire

vs. other approach to data collection to be an important issue, it is beyond our resources to seek and document evidence of all those circumstances in which a questionnaire may or may not be suitable. We will therefore simply highlight that this is an issue to be considered at the beginning of any data collection exercise and advise that an initial step should be a literature review to assess the suitability or otherwise of a questionnaire approach in the particular context of the study in hand.

- Studies of the applicability of the questionnaire approach (i.e. whether questionnaires are a feasible method of data collection) to particular subject groups (e.g. elderly, homeless or mentally ill people) should be excluded. It is similarly beyond our resources to seek and document evidence of the applicability or otherwise of the questionnaire approach across all possible subject groups. Once again, we will simply highlight that this is an issue to be considered at the beginning of any data collection exercise and advise that an initial step should be a literature review to assess the suitability or otherwise of a questionnaire approach in the particular context of the study in hand.
- Studies that simply report on the **use** of questionnaires/interviews should be excluded.
- Studies focusing on the use of interviews/questionnaires for clinical purposes (e.g. screening, health history taking, counselling) should be excluded. This means that papers reporting on methods for training clinical interviewers or improving their performance should also be excluded.
- Job/selection interviews should be excluded.
- Media interviews should be excluded.
- Studies comparing interview/questionnaire approaches with objective assessments (e.g. clinical examination, record-based approaches) should be excluded. Once again, this goes back to the issue of whether a questionnaire is a valid and appropriate means of data collection in a particular context, or whether an objective assessment is required. Our initial scan of the literature suggests that this is context and topic specific. As before, we will highlight this as an issue for consideration at the beginning of any data collection exercise and advise that an initial step should be a literature review to assess the suitability or otherwise of a questionnaire approach in the particular context of the study in hand.
- Studies addressing the development or refinement of a specific health status/quality of life measure should be excluded. This is already fairly well covered in the literature. Some of the other funded reviews in the HTA Methodology initiative, and studies funded from other sources, are addressing certain aspects of this.
- Studies (whether comparative or not) assessing the practical or psychometric properties of specific health status/quality of life measures should be excluded. Again, this is already fairly well covered in the literature. Some of the other funded reviews in the HTA Methodology Programme are addressing certain aspects of this.
- Comparison of short form vs. long form measures should be excluded if the focus of the paper is solely on the psychometric properties (validity and reliability) of the two versions. We believe that this information is likely to be context and measure specific and would not therefore add to the overall body of knowledge on best practice. However, we will note that altering the length of a measure may have an impact on reliability and validity, as well as on response rates and other factors, and that this needs to be borne in mind in choosing and using a measure.
- Comparison of proxies vs. self-report should be excluded if the focus of the paper is solely on the psychometric properties (validity and reliability) of the two approaches. Again, this information is likely to be context and measure specific and would not therefore add to the overall body of knowledge on best practice.
- Studies of research ethics in general terms should be excluded; ethical issues form the subject matter of another review funded under the HTA initiative.
- Papers solely reporting on the context in which telephone interviews are employed should be excluded; telephone interviews are not a novel topic.
- Papers solely reporting on the context in which non-interactive computerised techniques are employed should be excluded, again on the grounds that this is no longer a novel topic.
- Papers reporting techniques for establishing reliability, validity or responsiveness to change should be excluded; this is not a review of psychometric principles and methods.
- Papers reporting on the impact of questionnaires on respondents (e.g. whether they generate anxiety) should be excluded; such information is likely to be context or topic specific and general messages on "best practice" are not likely to be available.
- *Post-hoc* comparisons of respondents and non-respondents to specific surveys should be excluded. Although there may be some general trends with respect to who does and who does not respond, this information is not likely to

inform best practice. However, outputs from the review should mention the likelihood of non-response bias and remind people to anticipate it and test for it.
- The use of census data in coding occupations should be excluded; this is a highly specialised topic and unlikely to be of common concern to the target audience.
- Data checking and cleaning techniques should be excluded. This is likely to be highly context specific and specific to the software packages used for data entry, validation and analysis.
- Sampling methods should be excluded; other groups(s) within the HTA initiative will be funded to review sampling issues.

Language

Owing to resource constraints, we will include only studies written in the English language.

Time-scale

Owing to resource constraints, a time cut-off is also required. Many of the classic texts on questionnaire design and survey methodology date from the mid-1970s. That period also marked the growth of interest in the use of patient and staff questionnaires in outcome assessment. We will therefore take 1975 as an initial cut-off point for inclusion. If there is insufficient evidence from 1975 or later, evidence from earlier studies will be sought.

Study designs

Code	Type of design	Notes
1 Randomised (experimental) designs		
1	Randomised controlled trials	Participants are allocated to an intervention or a concurrent control group in a random manner (e.g. using random numbers tables, sealed envelopes, computerised systems). Factorial designs and cross-over designs are particular types of randomised controlled trial. Also included in this category are designs in which the entire sample is selected by some random process (e.g. random numbers, random telephone digit dialling) and selected participants are subsequently allocated to intervention and control groups in a systematic manner (e.g. alternation).
2 Non-randomised (quasi-experimental) designs		
2a	Non-random concurrent controlled studies	Participants are not randomly assigned to groups but intervention groups and control groups are concurrently assessed (e.g. a design in which participants at site A receive version 1 of a questionnaire, while those at site B receive version 2; sites A and B should be comparable at baseline). The manipulation of assignment of participants to groups (e.g. sites, interviewers) should be deliberate and carried out at the outset of the study; studies that simply involve a retrospective comparison of respondents allocated to different groups should be coded as cross-sectional studies (3b). However, manipulated studies where non-random (e.g. systematic, "every third patient") methods were used to allocate participants to intervention or control groups should be coded under this heading of non-random controlled studies.
2b	Self-controlled studies	These studies involve pre- and postintervention measures and are also termed before-and-after or longitudinal designs. Self-controlled studies can be strengthened by having a control group, also assessed at the same time points, who do not experience the intervention (a controlled before-and-after design), or by having more than one measurement point before and after the intervention (an interrupted time series design).

continued

contd

Code	Type of design	Notes
2c	Historically controlled studies	These studies make use of data collected for participants in other surveys. The most obvious example would be the case of a regularly repeated survey (e.g. Census, Labour Force Survey) where a change in methodology (e.g. question wording) was introduced at a specific point in time and the findings from the old version and the new version were compared. To be classed as a historically controlled study, the change must be deliberately manipulated (e.g. an explicit decision to change the question wording). If the change is not clearly deliberately manipulated, code as a cohort study (3b).
3 Descriptive (observational) designs		
3a	Cross-sectional studies	These are single-approach studies that provide descriptive data at one fixed point in time. An example might be a description of the conduct and outcome of an interactive computer-assisted interview study. Studies involving a purely retrospective comparison of participants allocated to groups (e.g. sites, interviewers) should also be included in this category.
3b	Cohort studies	These are prospective studies that provide information about a specific group and changes therein. There are two main types of cohort study. The first focuses on the same population each time, but the samples may be different (e.g. 5-year follow-up surveys of people who graduated in 1980, with a separate sample being selected each time). The second focuses on the same sample each time and is often termed a panel survey. Studies of recurrent surveys (e.g. Census, Labour Force Survey) with an opportunistic, rather than a deliberately manipulated, change at a specific time point should be coded as cohort studies.
3c	Case-control studies	These are retrospective studies used to try to explain a current phenomenon. They involve at least two groups. Cases are those who exhibit the phenomenon of interest, controls those who do not. A retrospective comparison of responders and non-responders to a particular survey would be an example of a case-control study; we have explicitly excluded such *post-hoc* comparisons of respondents and non-respondents in our protocol (see appendix 1). Other examples of case-control studies in the context of this review are difficult to identify and are unlikely to occur.
4 Review papers		
4a	Meta-analysis based on systematic review	Explicit criteria are used for the identification and inclusion of papers. Estimates of the treatment effect from identified studies are quantitatively pooled or combined.
4b	Systematic review without meta-analysis	Explicit criteria are used for identification and inclusion of papers. Estimates of the treatment effect are either combined qualitatively or not combined at all.
4c	Meta-analysis based on non-systematic review	The methods and criteria used to identify and include papers are not explicitly stated. Estimates of the treatment effect from identified studies are quantitatively pooled or combined.
4d	Non-systematic review without meta-analysis	The methods and criteria used to identify and include papers are not explicitly stated. Estimates of the treatment effect are either combined qualitatively or not combined at all.

continued

contd

Code	Type of design	Notes
5	***Theoretical papers***	
5a	Theoretical papers	A theory about behaviour relevant to survey methods is presented. An example would be a paper on the psychology of perception, such as the way in which people scan a page.
5b	Position papers	An advocated methodology is presented. An example would be a paper on Dillman's Total Survey Method approach. If the paper includes a "case study" of an actual survey, code according to the type of descriptive study (most likely to be a cross-sectional study, code 3a).

Notes:

1. The terms "split half" and "split ballot" are frequently used in survey methods research. Although many split-half and split-ballot designs are randomised controlled trials, this may not always be the case; some may split the groups systematically rather than randomly. Check carefully how the split was made before allocating a code.
2. The Solomon Four-Group design is a particular sort of randomised controlled trial, involving, as the name suggests, four groups. Groups 1 and 2 receive a pre-intervention measurement; groups 3 and 4 do not. The intervention is applied to groups 1 and 3, while groups 2 and 4 act as controls. All four groups receive a postintervention measurement. Such a design should be coded simply as a randomised controlled trial.
3. Similarly, the various sorts of factorial design (e.g. Latin Square, $2 \times 2 \times 2$ factorial design) used in the context of randomised controlled trials, should simply be coded as randomised controlled trials.

Appendix 5

Additional studies

As outlined on page 13, owing to human error, identified references from MEDLINE for 1987–1992, and from PsycLIT for 1979, 1991 and 1993–1996, were inadvertently omitted from the original Reference Manager database. On applying the search strategies described on page 12 to these missing years in the respective electronic databases, a further 344 articles were identified. An initial sift on the basis of title and abstract, as described on page 12, indicated that 220 of these did not meet the inclusion criteria. Hard copies of the remaining 124 articles were sought and evaluated against the inclusion criteria and the 5-point methodological screen (described on page 13 and in appendix 3). On further screening, 46 articles were rejected as being out of scope or not meeting minimum methodological criteria. The 78 articles that were deemed to be in scope and passed the 5-point methodological screen were categorised according to the main sections and subsections of this review (chapters 3–6). These are listed below. Some articles involved manipulation of more than one aspect of the survey process and are therefore listed under multiple headings. The quality scoring system described on page 14 was not applied to these additional articles; neither was any attempt made to abstract data from them. Because of the deadline for the completion of this project, it was not possible retrospectively to incorporate the findings from these articles.

Self-completion questionnaires versus telephone interviews

1. Aquilino WS. Interview mode effects in surveys of drug and alcohol use: a field experiment. *Public Opinion Quarterly* 1994;**58**:210–40.
2. Dillman DA, West KK, Clark JR. Influence of an invitation to answer by telephone on response to census questionnaires. *Public Opinion Quarterly* 1994;**58**:557–68.
3. Fournier L, Kovess V. A comparison of mail and tele-phone interview strategies for mental health surveys. *Canadian Journal of Psychiatry* 1993;**38**:525–33.
4. Zapka JG, Chasan-Taber L, Bigelow C, Hurley T. Methodological issues for health-related surveys of multicultural older women. *Evaluation and the Health Professions* 1994;**17**:485–500.

Self-completion questionnaires versus face-to-face interviews

1. Aquilino WS. Interview mode effects in surveys of drug and alcohol use: a field experiment. *Public Opinion Quarterly* 1994;**58**:210–40.
2. Bishop GF, Fisher BS. "Secret ballots" and self-reports in an exit-poll experiment. *Public Opinion Quarterly* 1995;**59**:568–88.
3. Bush AJ, Bush RF, Chen HC. Method of administration effects in mall intercept interviews. *Journal of the Market Research Society* 1991;**33**:309–19.
4. Doll H, McPherson K, Davies J, Flood A, Smith J, Williams G *et al.* Reliability of questionnaire responses as compared with interview in the elderly: views of the outcome of transurethral resection of the prostate. *Social Science and Medicine* 1991;**33**:1303–8.
5. Jackson N, Little J, Wilson AD. Comparison of diet history interview and self-completed questionnaire in assessment of diet in an elderly population. *Journal of Epidemiology and Community Health* 1990;**44**:162–9.
6. Krysan M, Schuman H, Scott LJ, Beatty P. Response rates and response content in mail versus face-to-face surveys. *Public Opinion Quarterly* 1994;**58**:381–99.
7. Locke SD, Gilbert BO. Method of psychological assessment, self-disclosure, and experiential differences: a study of computer, questionnaire, and interview assessment formats. *Journal of Social Behavior and Personality* 1995;**10**:255–63.
8. Rolnick SJ, Gross CR, Garrard J, Gibson RW. A comparison of response rate, data quality, and cost in the collection of data on sexual history and personal behaviors. Mail survey approaches and in-person interview. *American Journal of Epidemiology* 1989;**129**:1052–61.
9. Sobell J, Block G, Koslowe P, Tobin J, Andres R. Validation of a retrospective questionnaire assessing diet 10–15 years ago. *American Journal of Epidemiology* 1989;**130**:173–87.
10. Tourangeau R, Smith TW. Asking sensitive questions: the impact of data collection mode, question format, and question context. *Public Opinion Quarterly* 1996;**60**:275–304.
11. Williams BL, Suen H. A methodological comparison of survey techniques in obtaining self-reports of condom-related behaviors. *Psychological Reports* 1994;**75**:1531–7.

Telephone interviews versus face-to-face interviews

1. Aquilino WS. Interview mode effects in surveys of drug and alcohol use: a field experiment. *Public Opinion Quarterly* 1994;**58**:210–40.
2. Chwalow AJ, Costagliola D, Stern J, Mesbah M, Eschwege E. Telephone versus face to face interviewing as a means of collecting data relevant to the management of diabetes among general practitioners in France: a randomized design. *Diabete et Metabolisme* 1989;**15**:157–60.
3. Gonzalez GM, Costello CR, Valenzuela M, Chaidez B, Nunez-Alvarez A. Bilingual computerized speech-recognition screening for clinical depression: evaluating a cellular telephone prototype. *Behavior Research Methods, Instruments and Computers* 1995;**27**:476–82.
4. Johnson TP, Hougland JG, Moore RW. Sex differences in reporting sensitive behavior: a comparison of interview methods. *Sex Roles* 1991;**24**:669–80.
5. Kaplan CP, Tanjasiri SP. The effects of interview mode on smoking attitudes and behavior: self-report among female Latino adolescents. *Substance Use and Misuse* 1996;**31**:947–63.
6. Korner-Bitensky N, Wood-Dauphinee S, Shapiro S, Becker R. A telephone interview compared to a face-to-face interview in determining health status of patients discharged home from a rehabilitation hospital. *Canadian Journal of Rehabilitation* 1993;**7**:73–5.

Computer-assisted versus paper-based self-completion questionnaires

1. Bratton GR, Newsted PR. Response effects and computer-administered questionnaires: the role of the entry task and previous computer experience. *Behaviour and Information Technology* 1995;**14**:300–12.
2. DiLalla DL. Computerized administration of the Multidimensional Personality Questionnaire. *Assessment* 1996;**3**:365–74.
3. Locke SD, Gilbert BO. Method of psychological assessment, self-disclosure, and experiential differences: a study of computer, questionnaire, and interview assessment formats. *Journal of Social Behavior and Personality* 1995;**10**:255–63.
4. Rosenfeld P, Booth-Kewley S, Edwards JE, Thomas MD. Responses on computer surveys: impression management, social desirability, and the Big Brother syndrome. *Computers in Human Behavior* 1996;**12**:263–74.

Other aspects of mode of administration

1. Bratton GR, Newsted PR. Response effects and computer-administered questionnaires: the role of the entry task and previous computer experience. *Behaviour and Information Technology* 1995;**14**:300–12.
2. Grossarth-Maticek R, Eysenck HJ, Barrett P. Prediction of cancer and coronary heart disease as a function of method of questionnaire administration. *Psychological Reports* 1993;**73**:943–59.
3. Kittleson MJ. An assessment of the response rate via the postal service and e-mail. *Health Values: The Journal of Health Behavior, Education and Promotion* 1995;**19**:27–39.
4. Tourangeau R, Smith TW. Asking sensitive questions: the impact of data collection mode, question format, and question context. *Public Opinion Quarterly* 1996;**60**:275–304.
5. Williams BL, Suen H. A methodological comparison of survey techniques in obtaining self-reports of condom-related behaviors. *Psychological Reports* 1994;**75**:1531–7.

Question wording

1. Abramson PR, Ostrom CW. Question wording and partisanship: change and continuity in party loyalties during the 1992 election campaign. *Public Opinion Quarterly* 1994;**58**:21–48.
2. Blair EA, Ganesh GK. Characteristics of interval-based estimates of autobiographical frequencies. *Applied Cognitive Psychology* 1991;**5**:237–50.
3. Britt MA. General versus elaborated questions in an employee opinion survey. *Journal of Social Behavior and Personality* 1993;**8**:335–40.
4. Burton S, Blair E. Task conditions, response formulation processes, and response accuracy for behavioral frequency questions in surveys. *Public Opinion Quarterly* 1991;**55**:50–79.
5. Schriesheim CA, Eisenbach RJ, Hill KD. The effect of negation and polar opposite item reversals on questionnaire reliability and validity: an experimental investigation. *Educational and Psychological Measurement* 1991;**51**:67–78.
6. Uitenbroek DG, McQueen DV. Leisure time physical activity in Scotland: trends 1987–1991 and the effect of question wording. *Sozial- und Praventivmedizin.* 1992;**37**:113–17.
7. Waenke M, Schwarz N, Noelle-Neumann E. Asking comparative questions: the impact of the direction of comparison. *Public Opinion Quarterly* 1995;**59**:347–72.

8. Wiederman MW, Weis DL, Allgeier ER. The effect of question preface on response rates to a telephone survey of sexual experience. *Archives of Sexual Behavior* 1994;**23**:203–15.

Question sequencing

1. Barnes JH, Banahan BF, Fish KE. The response effect of question order in computer-administered questioning in the social sciences. *Social Science Computer Review* 1995;**13**:47–53.

2. Benton JE, Daly JL. A question order effect in a local government survey. *Public Opinion Quarterly* 1991;**55**:640–2.

3. Bickart BA. Carryover and backfire effects in marketing research. *Journal of Marketing Research* 1993;**30**:52–62.

4. Frey JH. The impact of cover design and first questions on response rates for a mail survey of skydivers. *Leisure Sciences* 1991;**13**:67–76.

5. Hays RD, Bell RM, Hill LL, Gillogly JJ, Lewis MW, Marshall GN *et al.* The impact of response options and location in a microcomputer interview on drinking drivers' alcohol use self-reports. *Alcohol and Alcoholism* 1994;**29**:203–9.

6. King AC. Enhancing the self-report of alcohol consumption in the community: two questionnaire formats. *American Journal of Public Health* 1994;**84**:294–6.

7. Mason R, Carlson JE, Tourangeau R. Contrast effects and subtraction in part-whole questions. *Public Opinion Quarterly* 1994;**58**:569–78.

8. Melnick SA. The effects of item grouping on the reliability and scale scores of an affective measure. *Educational and Psychological Measurement* 1993;**53**:211–16.

9. Pourjalali H, Kimbrell J. Effects of four instrumental variables on survey response. *Psychological Reports* 1994;**75**:895–8.

10. Schwarz N, Hippler H-J. Subsequent questions may influence answers to preceding questions in mail surveys. *Public Opinion Quarterly* 1995;**59**:93–7.

11. Sheeran P, Orbell S. How confidently can we infer health beliefs from questionnaire responses? *Psychology and Health* 1996;**11**:273–90.

12. Tourangeau R, Smith TW. Asking sensitive questions: the impact of data collection mode, question format, and question context. *Public Opinion Quarterly* 1996;**60**:275–304.

13. Willits FK, Ke B. Part-whole question order effects: views of rurality. *Public Opinion Quarterly* 1995;**59**:392–403.

Response format

1. Burton S, Blair E. Task conditions, response formulation processes, and response accuracy for behavioral frequency questions in surveys. *Public Opinion Quarterly* 1991;**55**:50–79.

2. Gaskell GD, O'Muircheartaigh CA, Wright DB. Survey questions about the frequency of vaguely defined events: the effects of response alternatives. *Public Opinion Quarterly* 1994;**58**:241–54.

3. Hall T, Shelby B, Rolloff D. Effect of varied question format on boaters' norms. *Leisure Sciences* 1996;**18**:193–204.

4. Hays RD, Bell RM, Hill LL, Gillogly JJ, Lewis MW, Marshall GN *et al.* The impact of response options and location in a microcomputer interview on drinking drivers' alcohol use self-reports. *Alcohol and Alcoholism* 1994;**29**:203–9

5. Hunt DM, Magruder S, Bolon DS. Questionnaire format bias: when are juxtaposed scales appropriate: a call for further research. *Psychological Reports* 1995;**77**:931–41.

6. Larson CO, Hays RD, Nelson EC. Do the pictures influence scores on the Dartmouth COOP Charts? *Quality of Life Research* 1992;**1**:247–9.

7. Pfennings L, Cohen L, van der Ploeg H. Preconditions for sensitivity in measuring change: visual analogue scales compared to rating scales in a Likert format. *Psychological Reports* 1995;**77**:475–80.

8. Rasinski KA, Mingay D, Bradburn NM. Do respondents really "mark all that apply" on self-administered questions? *Public Opinion Quarterly* 1994;**58**:400–8.

9. Sekely WS, Blakney VL. The effect of response position on trade magazine readership and usage. *Journal of Advertising Research* 1994;**34**:53–60.

10. Tourangeau R, Smith TW. Asking sensitive questions: the impact of data collection mode, question format, and question context. *Public Opinion Quarterly* 1996;**60**:275–304.

11. Waenke M, Schwarz N, Noelle-Neumann E. Asking comparative questions: the impact of the direction of comparison. *Public Opinion Quarterly* 1995;**59**:347–72.

Length of questionnaire

1. Biner PM, Kidd HJ. The interactive effects of monetary incentive justification and questionnaire length on mail survey response rates. *Psychology and Marketing* 1994;**11**:483–92.

2. Childers TL, Ferrell OC. Response rates and perceived questionnaire length in mail surveys. *Journal of Marketing Research* 1979;**16**:429–31.

3. Dillman DA, Sinclair MD, Clark JR. Effects of questionnaire length, respondent-friendly design, and a difficult question on response rates for occupant-addressed census mail surveys. *Public Opinion Quarterly* 1993;**57**:289–304.

4. Lerman Y, Slepon R, Kark JD. The effect of using short versus detailed self-administered questionnaires on the estimate of illicit drug use among young adults. *Drug and Alcohol Review* 1995;**14**:377–84.

5. Rolnick SJ, Gross CR, Garrard J, Gibson RW. A comparison of response rate, data quality, and cost in the collection of data on sexual history and personal behaviors. Mail survey approaches and in-person interview. *American Journal of Epidemiology* 1989;**129**:1052–61.

6. Rucker MH, Arbaugh JE. A comparison of matrix questionnaires with standard questionnaires. *Educational and Psychological Measurement* 1979;**39**:637–43.

Pagination

1. Childers TL, Ferrell OC. Response rates and perceived questionnaire length in mail surveys. *Journal of Marketing Research* 1979;**16**:429–31.

2. Dillman DA, Sinclair MD, Clark JR. Effects of questionnaire length, respondent-friendly design, and a difficult question on response rates for occupant-addressed census mail surveys. *Public Opinion Quarterly* 1993;**57**:289–304.

Paper colour and quality

No further studies identified.

Print details

No further studies identified.

Cover design

1. Frey JH. The impact of cover design and first questions on response rates for a mail survey of skydivers. *Leisure Sciences* 1991;**13**:67–76.

Question and response category format

1. Kelly A, Knaap G, Simon A, Temperley S. The effects of "preprinting" on survey and item response rates: a research note. *Journal of Leisure Research* 1996;**28**:122–8.

2. Pfennings L, Cohen L, van der Ploeg H. Preconditions for sensitivity in measuring change: visual analogue scales compared to rating scales in a Likert format. *Psychological Reports* 1995;**77**:475–80.

3. Rasinski KA, Mingay D, Bradburn NM. Do respondents really "mark all that apply" on self-administered questions? *Public Opinion Quarterly* 1994;**58**:400–8.

4. Rucker MH, Arbaugh JE. A comparison of matrix questionnaires with standard questionnaires. *Educational and Psychological Measurement* 1979;**39**:637–43.

Instructions

1. Willits FK, Ke B. Part-whole question order effects: views of rurality. *Public Opinion Quarterly* 1995;**59**:392–403.

Timing of survey

No further studies identified.

Number and relative timing of contacts

No further studies identified.

Prenotification contacts

1. Chebat JC, Picard J. Does prenotification increase response rates in mail surveys? A self-perception approach. *Journal of Social Psychology* 1991;**131**:477–81.

2. Childers TL, Skinner SJ. Gaining respondent cooperation in mail surveys through prior commitment. *Public Opinion Quarterly* 1979;**43**:558–61.

3. Furst LG, Blitchington WP. The use of a descriptive cover letter and secretary pre-letter to increase response rate in a mailed survey. *Personnel Psychology* 1979;**32**:155–9.

Follow-up contacts (reminders)

1. Chapman S, Wong WL. Incentives for questionnaire respondents. *Australian Journal of Public Health* 1991;**15**:66–7.

2. Dillman DA, West KK, Clark JR. Influence of an invitation to answer by telephone on response to census questionnaires. *Public Opinion Quarterly* 1994;**58**:557–68.

3. Maheux B, Legault C, Lambert J. Increasing response rates in physicians' mail surveys: an experimental study. *American Journal of Public Health* 1989;**79**:638–9.

Postal rates and types

1. Choi BC, Pak AW, Purdham JT. Effects of mailing strategies on response rate, response time, and cost in a questionnaire study among nurses. *Epidemiology* 1990;**1**:72–4.

2. Gitelson R, Kerstetter D, Guadagnolo F. Research note: the impact of incentives and three forms of postage on mail survey response rates. *Leisure Sciences* 1993;**15**:321–7.

3. Maheux B, Legault C, Lambert J. Increasing response rates in physicians' mail surveys: an experimental study. *American Journal of Public Health* 1989;**79**:638–9.

4. Moss VD, Worthen BR. Do personalization and postage make a difference on response rates to surveys of professional populations? *Psychological Reports* 1991;**68**:692–4.

5. Price JH, Easton A, Kandakai T, Oden L. Race-specific versus general stamps on African–American women's survey return rates. *Perceptual and Motor Skills* 2001;**82**:928–30.

Anonymity and confidentiality

1. Dillman DA, Singer E, Clark JR, Treat JB. Effects of benefits appeals, mandatory appeals, and variations in statements of confidentiality on completion rates for census questionnaires. *Public Opinion Quarterly* 1996;**60**:376–89.

2. Kalafatis SP, Blankson C. An investigation into the effect of questionnaire identification numbers in consumer mail surveys. *Journal of the Market Research Society* 1996;**38**:277–84.

Personalisation

1. Maheux B, Legault C, Lambert J. Increasing response rates in physicians' mail surveys: an experimental study. *American Journal of Public Health* 1989;**79**:638–9.

2. Moss VD, Worthen BR. Do personalization and postage make a difference on response rates to surveys of professional populations? *Psychological Reports* 1991;**68**:692–4.

Covering letters

1. Biner PM, Kidd HJ. The interactive effects of monetary incentive justification and questionnaire length on mail survey response rates. *Psychology and Marketing* 1994;**11**:483–92.

2. Camunas C, Alward RR, Vecchione E. Survey response rates to a professional association mail questionnaire. *Journal of the New York State Nurses Association* 1990;**21**:7–9.

3. Dillman DA, Singer E, Clark JR, Treat JB. Effects of benefits appeals, mandatory appeals, and variations in statements of confidentiality on completion rates for census questionnaires. *Public Opinion Quarterly* 1996;**60**:376–89.

4. Furst LG, Blitchington WP. The use of a descriptive cover letter and secretary pre-letter to increase response rate in a mailed survey. *Personnel Psychology* 1979;**32**:155–9.

5. Gendall P, Hoek J, Esslemont D. The effect of appeal, complexity and tone in a mail survey covering letter. *Journal of the Market Research Society* 1995;**37**:251–68.

6. Maheux B, Legault C, Lambert J. Increasing response rates in physicians' mail surveys: an experimental study. *American Journal of Public Health* 1989;**79**:638–9.

7. Pourjalali H, Kimbrell J. Effects of four instrumental variables on survey response. *Psychological Reports* 1994;**75**:895–8.

8. Shaker LA, Derycke-Chapman K, Brass LM. Using pamphlets with mail surveys to improve response. *American Journal of Public Health* 1992;**82**:463–4.

Sponsorship

1. Etter J-F, Perneger TV, Rougemont A. Does sponsorship matter in patient satisfaction surveys? A randomized trial. *Medical Care* 1996;**34**:327–35.

Saliency/subject matter

1. Barker PJ, Cooper RF. Do sexual health questions alter the public's response to lifestyle questionnaires? *Journal of Epidemiology and Community Health* 1996;**50**:688.

2. Dillman DA, Sinclair MD, Clark JR. Effects of questionnaire length, respondent-friendly design, and a difficult question on response rates for occupant-addressed census mail surveys. *Public Opinion Quarterly* 1993;**57**:289–304.

3. Martin CL. The impact of topic interest on mail survey response behaviour. *Journal of the Market Research Society* 1994;**36**:327–38.

4. Windsor J. What can you ask about? The effect on response to a postal screen of asking about two potentially sensitive questions. *Journal of Epidemiology and Community Health* 1992;**46**:83–5.

Incentives

1. Biner PM, Kidd HJ. The interactive effects of monetary incentive justification and questionnaire length on mail survey response rates. *Psychology and Marketing* 1994;**11**:483–92.
2. Brennan M, Hoek J, Astridge C. The effects of monetary incentives on the response rate and cost-effectiveness of a mail survey. *Journal of the Market Research Society* 1991;**33**:229–41.
3. Camunas C, Alward RR, Vecchione E. Survey response rates to a professional association mail questionnaire. *Journal of the New York State Nurses Association* 1990;**21**:7–9.
4. Chapman S, Wong WL. Incentives for questionnaire respondents. *Australian Journal of Public Health* 1991;**15**:66–7.
5. Chebat JC, Picard J. Does prenotification increase response rates in mail surveys? A self-perception approach. *Journal of Social Psychology* 1991;**131**:477–81.
6. Gitelson R, Kerstetter D, Guadagnolo F. Research note: the impact of incentives and three forms of postage on mail survey response rates. *Leisure Sciences* 1993;**15**:321–7.
7. Kalafatis SP, Madden FJ. The effect of discount coupons and gifts on mail survey response rates among high involvement respondents. *Journal of the Market Research Society* 1995;**37**:171–84.
8. Marrett LD, Kreiger N, Dodds L, Hilditch S. The effect on response rates of offering a small incentive with a mailed questionnaire. *Annals of Epidemiology* 1992;**2**:745–3.

Feedback of results

No further studies identified.

Miscellaneous

No further studies identified.

Methodology Group

Members

Methodology Programme Director
Professor Richard Lilford
Director of Research and Development
NHS Executive – West Midlands, Birmingham

Chair
Professor Martin Buxton
Director, Health Economics Research Group
Brunel University, Uxbridge

Professor Douglas Altman
Professor of Statistics in Medicine
University of Oxford

Dr David Armstrong
Reader in Sociology as Applied to Medicine
King's College, London

Professor Nicholas Black
Professor of Health Services Research
London School of Hygiene & Tropical Medicine

Professor Ann Bowling
Professor of Health Services Research
University College London Medical School

Professor David Chadwick
Professor of Neurology
The Walton Centre for Neurology & Neurosurgery
Liverpool

Dr Mike Clarke
Associate Director (Research)
UK Cochrane Centre, Oxford

Professor Paul Dieppe
Director, MRC Health Services Research Centre
University of Bristol

Professor Michael Drummond
Director, Centre for Health Economics
University of York

Dr Vikki Entwistle
Senior Research Fellow, Health Services Research Unit
University of Aberdeen

Professor Ewan B Ferlie
Professor of Public Services Management
Imperial College, London

Professor Ray Fitzpatrick
Professor of Public Health & Primary Care
University of Oxford

Dr Naomi Fulop
Deputy Director, Service Delivery & Organisation Programme
London School of Hygiene & Tropical Medicine

Mrs Jenny Griffin
Head, Policy Research Programme
Department of Health
London

Professor Jeremy Grimshaw
Programme Director
Health Services Research Unit
University of Aberdeen

Professor Stephen Harrison
Professor of Social Policy
University of Manchester

Mr John Henderson
Economic Advisor
Department of Health, London

Professor Theresa Marteau
Director, Psychology & Genetics Research Group
Guy's, King's & St Thomas's School of Medicine, London

Dr Henry McQuay
Clinical Reader in Pain Relief
University of Oxford

Dr Nick Payne
Consultant Senior Lecturer in Public Health Medicine
ScHARR
University of Sheffield

Professor Joy Townsend
Director, Centre for Research in Primary & Community Care
University of Hertfordshire

Professor Kent Woods
Director, NHS HTA Programme, & Professor of Therapeutics
University of Leicester

continued

HTA Commissioning Board

Members

Programme Director
Professor Kent Woods
Director, NHS HTA Programme, &
Professor of Therapeutics
University of Leicester

Chair
Professor Shah Ebrahim
Professor of Epidemiology of Ageing
University of Bristol

Deputy Chair
Professor Jon Nicholl
Director, Medical Care Research Unit
University of Sheffield

Professor Douglas Altman
Director, ICRF Medical Statistics Group
University of Oxford

Professor John Bond
Director, Centre for Health Services Research
University of Newcastle-upon-Tyne

Ms Christine Clark
Freelance Medical Writer
Bury, Lancs

Professor Martin Eccles
Professor of Clinical Effectiveness
University of Newcastle-upon-Tyne

Dr Andrew Farmer
General Practitioner &
NHS R&D Clinical Scientist
Institute of Health Sciences
University of Oxford

Professor Adrian Grant
Director, Health Services Research Unit
University of Aberdeen

Dr Alastair Gray
Director, Health Economics Research Centre
Institute of Health Sciences
University of Oxford

Professor Mark Haggard
Director, MRC Institute of Hearing Research
University of Nottingham

Professor Jenny Hewison
Senior Lecturer
School of Psychology
University of Leeds

Professor Alison Kitson
Director, Royal College of Nursing Institute, London

Dr Donna Lamping
Head, Health Services Research Unit
London School of Hygiene & Tropical Medicine

Professor David Neal
Professor of Surgery
University of Newcastle-upon-Tyne

Professor Gillian Parker
Nuffield Professor of Community Care
University of Leicester

Dr Tim Peters
Reader in Medical Statistics
University of Bristol

Professor Martin Severs
Professor in Elderly Health Care
University of Portsmouth

Dr Sarah Stewart-Brown
Director, Health Services Research Unit
University of Oxford

Professor Ala Szczepura
Director, Centre for Health Services Studies
University of Warwick

Dr Gillian Vivian
Consultant in Nuclear Medicine & Radiology
Royal Cornwall Hospitals Trust
Truro

Professor Graham Watt
Department of General Practice
University of Glasgow

Dr Jeremy Wyatt
Senior Fellow
Health Knowledge Management Centre
University College London

Current and past membership details of all HTA 'committees' are available from the HTA website (see inside front cover for details)

Feedback

The HTA Programme and the authors would like to know your views about this report.

The Correspondence Page on the HTA website (http://www.ncchta.org) is a convenient way to publish your comments. If you prefer, you can send your comments to the address below, telling us whether you would like us to transfer them to the website.

We look forward to hearing from you.

Copies of this report can be obtained from:

The National Coordinating Centre for Health Technology Assessment,
Mailpoint 728, Boldrewood,
University of Southampton,
Southampton, SO16 7PX, UK.
Fax: +44 (0) 23 8059 5639 Email: hta@soton.ac.uk
http://www.ncchta.org

ISSN 1366-5278