Health Technology Assessment 2012; Vol. 16: No. 27
ISSN 1366-5278

Mode of data elicitation, acquisition and response to surveys: a systematic review

K Hood, M Robling, D Ingledew, D Gillespie, G Greene, R Ivins, I Russell, A Sayers, C Shaw and J Williams

May 2012
10.3310/hta16270

Health Technology Assessment
NIHR HTA programme
www.hta.ac.uk

How to obtain copies of this and other HTA programme reports

An electronic version of this title, in Adobe Acrobat format, is available for downloading free of charge for personal use from the HTA website (www.hta.ac.uk). A fully searchable DVD is also available (see below).

Printed copies of HTA journal series issues cost £20 each (post and packing free in the UK) to both public **and** private sector purchasers from our despatch agents.

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

How to order:

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

Additionally the HTA website allows you to either print out your order or download a blank order form.

Contact details are as follows:

Synergie UK (HTA Department)
Digital House, The Loddon Centre
Wade Road
Basingstoke
Hants RG24 8QW

Email: orders@hta.ac.uk

Tel: 0845 812 4000 – ask for ‘HTA Payment Services’
(out-of-hours answer-phone service)

Fax: 0845 812 4001 – put ‘HTA Order’ on the fax header

Payment methods

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

Paying by credit card
You can order using your credit card by phone, fax or post.

Subscriptions

NHS libraries can subscribe free of charge. Public libraries can subscribe at a reduced cost of £100 for each volume (normally comprising 40–50 titles). The commercial subscription rate is £400 per volume (addresses within the UK) and £600 per volume (addresses outside the UK). Please see our website for details. Subscriptions can be purchased only for the current or forthcoming volume.

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

Please use the form on the HTA website (www.hta.ac.uk/htacd/index.shtml). *HTA on DVD* is currently free of charge worldwide.

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

Mode of data elicitation, acquisition and response to surveys: a systematic review

K Hood,[1*] M Robling,[1] D Ingledew,[2] D Gillespie,[1] G Greene,[1] R Ivins,[1] I Russell,[3] A Sayers,[4] C Shaw[5] and J Williams[3]

[1]School of Medicine, Cardiff University, Cardiff, UK
[2]School of Psychology, Bangor University, Bangor, UK
[3]School of Medicine, Swansea University, Swansea, UK
[4]Faculty of Medicine and Dentistry, University of Bristol, Bristol, UK
[5]Faculty of Health, Sport and Science, University of Glamorgan, Ponytypridd, UK

*Corresponding author

Declared competing interests of authors: none

Published May 2012
DOI: 10.3310/hta16270

This report should be referenced as follows:

Hood K, Robling M, Ingledew D, Gillespie D, Greene G, Ivins R, *et al*. Mode of data elicitation, acquisition and response to surveys: a systematic review. *Health Technol Assess* 2012;**16**(27).

Health Technology Assessment is indexed and abstracted in *Index Medicus*/MEDLINE, *Excerpta Medica*/EMBASE, *Science Citation Index Expanded* (*SciSearch*®) and *Current Contents*®/ Clinical Medicine.

NIHR Health Technology Assessment programme

The Health Technology Assessment (HTA) programme, part of the National Institute for Health Research (NIHR), was set up in 1993. It produces high-quality research information on the effectiveness, costs and broader impact of health technologies for those who use, manage and provide care in the NHS. 'Health technologies' are broadly defined as all interventions used to promote health, prevent and treat disease, and improve rehabilitation and long-term care.

The research findings from the HTA programme directly influence decision-making bodies such as the National Institute for Health and Clinical Excellence (NICE) and the National Screening Committee (NSC). HTA findings also help to improve the quality of clinical practice in the NHS indirectly in that they form a key component of the 'National Knowledge Service'.

The HTA programme is needs led in that it fills gaps in the evidence needed by the NHS. There are three routes to the start of projects.

First is the commissioned route. Suggestions for research are actively sought from people working in the NHS, from the public and consumer groups and from professional bodies such as royal colleges and NHS trusts. These suggestions are carefully prioritised by panels of independent experts (including NHS service users). The HTA programme then commissions the research by competitive tender.

Second, the HTA programme provides grants for clinical trials for researchers who identify research questions. These are assessed for importance to patients and the NHS, and scientific rigour.

Third, through its Technology Assessment Report (TAR) call-off contract, the HTA programme commissions bespoke reports, principally for NICE, but also for other policy-makers. TARs bring together evidence on the value of specific technologies.

Some HTA research projects, including TARs, may take only months, others need several years. They can cost from as little as £40,000 to over £1 million, and may involve synthesising existing evidence, undertaking a trial, or other research collecting new data to answer a research problem.

The final reports from HTA projects are peer reviewed by a number of independent expert referees before publication in the widely read journal series *Health Technology Assessment*.

Criteria for inclusion in the HTA journal series

Reports are published in the HTA journal series if (1) they have resulted from work 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 research reported in this issue of the journal was commissioned by the National Coordinating Centre for Research Methodology (NCCRM), and was formally transferred to the HTA programme in April 2007 under the newly established NIHR Methodology Panel. The HTA programme project number is 06/91/07. The contractual start date was in June 2005. The draft report began editorial review in September 2010 and was accepted for publication in March 2011. The commissioning brief was devised by the NCCRM who specified the research question and study design. The authors have been wholly responsible for all data collection, analysis and interpretation, and for writing up their work. The HTA editors and publisher have tried to ensure the accuracy of the authors' report and would like to thank the referees for their constructive comments on the draft document. However, they do not accept liability for damages or losses arising from material published in this report.

The views expressed in this publication are those of the authors and not necessarily those of the HTA programme or the Department of Health.

ISSN 1366-5278 (Print)
ISSN 2046-4924 (Online)
ISSN 2046-4932 (DVD)

Published by Prepress Projects Ltd, Perth, Scotland (www.prepress-projects.co.uk), on behalf of NETSCC, HTA.
Printed on acid-free paper in the UK by Charlesworth Press. MR

Abstract

Mode of data elicitation, acquisition and response to surveys: a systematic review

K Hood,[1]* M Robling,[1] D Ingledew,[2] D Gillespie,[1] G Greene,[1] R Ivins,[1] I Russell,[3] A Sayers,[4] C Shaw[5] and J Williams[3]

[1]School of Medicine, Cardiff University, Cardiff, UK
[2]School of Psychology, Bangor University, Bangor, UK
[3]School of Medicine, Swansea University, Swansea, UK
[4]Faculty of Medicine and Dentistry, University of Bristol, Bristol, UK
[5]Faculty of Health, Sport and Science, University of Glamorgan, Ponytypridd, UK

*Corresponding author

Background: Many studies in health sciences research rely on collecting participant-reported outcomes and attention is increasingly being paid to the mode of data collection. Consideration needs to be given to the validity of response via different modes and the impact that choice of mode might have on study conclusions.

Objectives: (1) To provide an overview of the theoretical models of survey response and how they relate to health research; (2) to review all studies comparing two modes of administration for subjective outcomes and assess the impact of mode of administration on response quality; (3) to explore the impact of findings for key identified health-related measures; and (4) to inform the analysis of multimode studies.

Data sources: A broad range of databases (for example EMBASE, PsychINFO, MEDLINE, EconLit, SPORTDiscus, etc.) were chosen to allow as comprehensive a selection as possible, and they were searched up until the end of 2004.

Review methods: The abstracts were reviewed against inclusion/exclusion criteria. Full papers were retrieved for all selected abstracts and then screened again using more detailed inclusion criteria related to the measures used. Papers that were still included were reviewed in full and detailed data extracted. At each stage, abstracts or papers were reviewed by a single reviewer.

Results: The search strategy identified 39,253 unique references, of which 2156 were considered as full papers, with 381 finally included in the review. Two features of mode were clearly associated with bias in response; however, none of the features of mode was associated with changes in precision. How the measure was administered, by an interviewer or by the person themselves, was highly significantly associated with bias ($p < 0.001$). A difference in sensory stimuli was also significant ($p = 0.03$). When both of these were present the average overall bias was < 1 point on a percentage scale. In terms of mediating factors, there was some suggestion that there was an interaction between both telephone and computer for data collection and date of publication, supporting the theory that differences disappear as new technologies become commonplace. Single-item measures were also related to greater degrees of bias than multi-item scales ($p = 0.01$). Individual analysis of the Short Form questionnaire-36 items and Minnesota Multiphasic Personality Inventory (MMPI) showed a varied pattern across the different subscales, with conflicting results between the two types of study. None of the MMPI measures used to detect deviant responding showed a relationship with the mode features tested. The limits of agreement analysis showed how variable measures were between modes at an individual rather than a group mean level.

Limitations: The search strategy covered the period up to 2004, so any new and emerging technologies were not included. Not all potential mode features were tested and there was limited information on potential mediating factors.

Conclusions: Researchers need to be aware of the different mode features that could have an impact on their results when selecting a mode of data collection for subjective outcomes. Further mode comparison studies, which manipulate mode features and directly assess impact over time, would be beneficial.

Funding: The National Institute for Health Research Health Technology Assessment programme.

Contents

Glossary

Acquiescence A response bias whereby respondents simply agree with an attitudinal statement regardless of content.

Optimising The process of carefully and comprehensively proceeding through all cognitive steps required when answering a survey question.

Satisficing A strategy of providing a satisfactory response to a survey question without the respondent expending the intended cognitive effort. This may be due to incomplete or biased or absent retrieval and/or integration of information when responding.

List of abbreviations

ACASI	audio computer-assisted self-interview
AUC	area under the curve
CAPI	computer-assisted personal interview
CASI	computer-assisted self-administered interview
CAT	computerised adaptive testing
CATI	computer-assisted telephone interview
CI	confidence interval
ES	effect size
HRQoL	health-related quality of life
ICC	intracluster correlation coefficient
IRT	item response theory
IVR	interactive voice response
MeSH	medical subject headings
MMPI	Minnesota Multiphasic Personality Inventory
PDA	personal digital assistant (handheld computer)
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROM	patient-reported outcome measure
QALY	quality-adjusted life-year
QoL	quality of life
RCT	randomised controlled trial
ROC	receiver operating characteristic
SAQ	self-administered questionnaire
SD	standard deviation
SF-36	Short Form questionnaire-36 items

All abbreviations that have been used in this report are listed here unless the abbreviation is well known (e.g. NHS), or it has been used only once, or it is a non-standard abbreviation used only in figures/tables/appendices, in which case the abbreviation is defined in the figure legend or in the notes at the end of the table.

Executive summary

Background

Many studies in health sciences research rely on collecting participant-reported outcomes. Although some of these are participant reports of factual information, such as adherence to drug regimes, that could be objectively validated, there is an increasing recognition of the importance of subjective measures such as attitude to, and perceptions of, health and services provision. Alongside the exponential increase in health-related literature devoted to participant-reported outcomes, attention is being paid to the method or mode of data collection. Much of this has been driven by the rapid development of new technologies, which can lead to increased ease, speed and efficiency of data capture alongside an increasing drive for maximising response rates. Survey methodologies (e.g. in the business, marketing, social and political sciences) have a literature base of their own, covering theory to practice, much of which has been only slowly recognised in the health arena. Few health-related outcome development papers indicate a theoretical approach to eliciting survey response and the focus in choosing a mode for a study is often based predominantly on improving response rates and minimising cost. The impact on the validity of response is not generally a consideration. In addition to this, in order to gain as complete a data set as possible, many studies are using multiple modes either to enhance participants' choice (e.g. opting for web- or paper-based surveys) or to improve follow-up rates (e.g. non-responders getting telephone data collection). Although for practical reasons these choices are entirely justifiable, consideration needs to be given to the validity of response via different modes and the impact that the choice of mode or modes might have on the conclusions from a study.

Objectives

- To provide an overview of the theoretical models of survey response and how they relate to health research.
- To review all studies comparing two modes of administration for subjective outcomes and assess the impact of mode of administration on response quality.
- To explore the impact of findings for key identified health-related measures.
- To create an accessible resource for health science researchers, which will advise on the impact of the selection of different modes of data collection on response.
- To inform the analysis of multimode studies.

Methods

In order to inform the systematic review of mode comparison studies, a review of the theoretical models and how they relate to the health domain was undertaken. This clarified the need to focus on features of mode rather than crude modes per se in order to understand the way in which responses to subjective outcomes could be affected. From this, a theoretical model based on Tourangeau was proposed with four main features: administration (interviewer or self), use of the telephone, use of the computer and sensory stimuli (audio, visual or both). Additional features were proposed that may belong in a model of response as well as potential mediating factors, such as cognitive challenge of questions. This approach was used to define the data extraction and coding classifications for studies.

Owing to the large body of literature relating to survey methodology which is published outside the health research arena, all studies that incorporate a mode comparison were included, regardless of setting. This led to a broad search strategy covering a wide range of disciplines. In order to target methodological studies, some innovations in search strategy that separate out the process from traditional reviews of the effectiveness of interventions were undertaken.

Identifying the literature

For a study to be included in the review it needed to:

1. provide evidence of a comparison between two modes of data collection of either the same question or the same set of questions referring to the same theoretical construct
2. compare a construct that is subjective and cannot be externally validated
3. explicitly reference a comparison in the analysis
4. collect quantitative data, i.e. use structured questions and answers.

Studies were excluded from the review if they involved:

1. a comparison between a quantitative measure and one or more qualitative data collection methods/analyses (e.g. unstructured interviews, focus groups)
2. a comparator derived from routine clinical records – unless explicit reference to specific self-reported construct is made within those records
3. a comparison between the response of two different judges, i.e. comparing a response from an individual to that made by someone other than the responder, for example a clinician providing a diagnosis.

A broad range of databases (for example EMBASE, PsychINFO, MEDLINE, EconLit, SPORTDiscus, etc.) were searched with no restrictions on start date or language. Searches were conducted up until the end of 2004. A matrix-based research strategy was developed and tested, searching for combinations of terms that would imply a mode comparison study.

Review process

The abstracts (and titles only for some foreign-language papers with no English abstract) were reviewed against the inclusion/exclusion criteria. Full papers were retrieved for all selected abstracts and then screened again using more detailed inclusion criteria related to the measures used. Papers that were still included were reviewed in full and detailed data extracted. At each stage, abstracts or papers were reviewed by a single reviewer after a period of training. Training for each stage included an assessment of reliability and sensitivity.

In order to assess the quality of the evidence contributing to this review, each paper was assessed for methodological quality. Assessing the quality of evidence becomes particularly challenging in reviews of studies having diverse methodologies. In this particular review, randomised controlled trials were not necessarily expected and so a more generic quality assessment tool was needed. A new tool was developed from two existing tools and tested.

Evidence synthesis

An overview of the studies identified is presented descriptively, highlighting the different mode features identified in the theory review. Those with appropriate data are subjected to quantitative methods of synthesis using exploratory metaregression to identify the association between mode features and differences in response. The primary analysis is based on three key summary statistics calculated for each comparison. These are the absolute difference between the means (standardised) of the two modes, the ratio of the largest to the smallest variance of the two modes and the effect size (ES; absolute mean difference/standard deviation) between two modes.

Between- and within-subject studies were analysed together, controlling for the study design. Analysis was conducted at two levels to account for clustering of comparisons within a study. This allowed for study-level characteristics, measure characteristics and mode features to be considered in a single model. The modelling approach assessed the four main mode features from the theoretical review, then tested the addition of other candidate features and then assessed model fit including other possible moderators of effect and identified interaction.

The two most frequently occurring outcomes – the Short Form questionnaire-36 items (SF-36) and the Minnesota Multiphasic Personality Inventory (MMPI) – are analysed in more depth using Mantel–Haenszel for between-group studies and Bland and Altman limits of agreements for within-group studies.

Results

The search strategy identified 39,253 unique references, of which 2156 were considered as full papers. Of these, 597 progressed to data extraction, with 381 finally included in the review. The most common reason (44%) for exclusion once the full paper was considered was that there was no actual mode comparison in the study. The majority of included studies were from North America (62%), with only 10% being from the UK.

Study designs were relatively evenly divided into between- and within-person studies (52% and 47%, respectively), with only 39% using some form of randomisation (random allocation for between-person studies and random ordering for within-person studies). In terms of quality assessment, most studies described their hypotheses and study design well, and drew appropriate conclusions (89%, 83% and 81% – good, respectively), but the description of participants, group allocation, potential impact of timing of data collection and presenting of variances was less good (22%, 50%, 27% and 35% – poor, respectively).

The 381 studies provided descriptions on 1282 outcome measures, of which 57% were health related. The most frequently reported outcomes were the SF-36 (17 studies) and the MMPI (9 studies). Thirty per cent of studies considered only a single outcome in their mode comparison, but most considered more (ranging from 1 to 21 outcomes). These studies also described a number of mode comparisons, giving in total 1522 comparisons between modes on multiple outcomes for analysis. Of these, 977 reported enough data to be included in the analysis of absolute mean differences, 910 in the analysis of the ratio of variances and 912 in the analysis of the ES.

Two features of mode were clearly associated with bias in response; however, none of the features of mode was associated with changes in precision. How the measure was administered, by an interviewer or by the person themselves, was highly significantly associated with bias ($p < 0.001$). A difference in sensory stimuli was also significant ($p = 0.03$). When both of these were present the average overall bias was < 1 point on a percentage scale. In terms of mediating factors, there was some suggestion that there was an interaction between both telephone and computer for data collection and date of publication, supporting the theory that differences disappear as new technologies become commonplace. Single-item measures were also related to greater degrees of bias than multi-item scales ($p = 0.01$).

Individual analysis of the SF-36 and MMPI showed a varied pattern across the different subscales, with conflicting results between the two types of study. None of the MMPI measures used to detect deviant responding showed a relationship with the mode features tested. The limits of agreement analysis showed how variable measures were between modes at an individual rather than at a group mean level.

Conclusions

Implications for researchers

Researchers need to be aware of the different mode features that could have an impact on their results when selecting a mode of data collection for subjective outcomes. If researchers use a mixture of modes within their study (commonly a change in mode if there is poor or non-response), then consideration needs to be given to ameliorating potential biases consequent on this and controlling for them in analysis.

The potential does exist for there to be simple correction factors developed; however, these are likely to be measure specific. In analysis of current mixed-mode studies, researchers cannot just assume that results are comparable where a difference in administration or sensory stimuli exists and they need either to undertake sensitivity analyses or to formally control for mode in the analysis.

Recommendations for future research (in priority order)

There are already numerous studies considering a large number of outcome measures. However, these need to be reported in a standardised way to allow researchers to be able to make informed decisions about choice of mode with a particular outcome in a population. The development of reporting standards akin to PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses), STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) or CONSORT (Consolidated Standards of Reporting Trials) for mode comparison studies is urgently needed and could build on the quality assessment tool developed here.

Further mode comparison studies are required, but these need to be experimentally designed to manipulate mode features and directly assess the impact. This is preferable to more studies comparing two modes at a relatively pragmatic level without consideration of those features. Studies need to give consideration to evaluation and direct testing of the impact of some of the mediators of mode effects, as the lack of data presented in papers in this review limited our ability to analyse this component.

Further primary studies need to be done to evaluate the impact of mode features over time. There was a suggestion across studies that this occurred for ‘new’ technologies for data collection (telephone and computer), but the ‘learning effect’ for any mode over time will be important to evaluate further in order to inform studies with long-term follow-up over multiple time points. The potential biasing impact of this ‘learning effect’ over time could be seen in single-mode studies as well as mixed-mode ones.

The focus of this review has been on measurement for research purposes and, therefore, has focused predominantly on the impact of mode features on estimated effects at a group level. However, the increasing use of subjective patient-reported outcomes in clinical practice means that considerable further work is required to consider measurement equivalence and reliability of assessment for individuals rather than groups.

Funding

Funding for this study was provided by the Health Technology Assessment programme of the National Institute for Health Research.

Chapter 1

Introduction

Many studies in health sciences research rely on collecting participant-reported outcomes of some form or another. Although some of these are participant reports of factual information, such as adherence to drug regimes, that could be objectively validated, there is increasing recognition of the importance of subjective measures, such as attitude to, and perceptions of, health and services provision. In addition to this, measures relating to health status which are not objectively measurable, such as quality of life (QoL), are becoming key secondary or even primary outcomes in many studies. This has led to a rapid growth in the development and validation of such measures. Few clinical trials, even with interventions pharmacological or surgical in nature, would be run today without measuring the patients' QoL and assessing the acceptability of the intervention being trialled. The US Food and Drug Administration has recognised the importance of the inclusion of such measures as QoL for registration purposes[1] and the National Institute for Health and Clinical Excellence incorporates quality-adjusted life-years (QALYs) as part of its decision-making process.

Alongside the exponential increase in health-related literature devoted to participant-reported outcomes (such as QoL), attention is being paid to the method or mode of data collection. Much of this has been driven by two main components: the rapid development of new technologies that can lead to increased ease, speed and efficiency of data capture, alongside an increasing drive for maximising response rates. This has led to a wide variety of options for mode of data collection being available to the health science researcher, with some studies adopting multiple approaches to follow up as many of the participants as possible. Although this approach may make sense pragmatically, it needs to be informed by an understanding of the participant's ability to respond and statistical adjustment for biases introduced by multimode usage.

Theoretical approach

Survey methodologies (e.g. in the business, marketing, social and political sciences) have a literature base of their own covering theory to practice, much of which has been only slowly recognised in the health arena. Few health-related outcome development papers indicate a theoretical approach to eliciting survey response.

Although theoretical approaches are rarely considered, there has been a focus on maximising data capture by improving response rates. Reviews have been conducted which consider how features of the survey instrument (e.g. presentation, length, incentives) impact on response rates.[2,3] There has also been an increase in ways in which such data are collected – the mode of data collection. With increasing levels of technology, a wider variety of modes are in use. The main focus in choosing a mode for a study appears to be based predominantly on improving response rates and minimising cost. The impact on the validity of response is not generally a consideration. In addition to this, in order to gain as complete a dataset as possible, many studies are using mixed modes either to enhance participants' choice (e.g. opting for web- or paper-based surveys) or to improve follow-up rates (i.e. non-responders getting telephone data collection). Although for practical reasons these choices are entirely justifiable, consideration needs to be given to the validity of response via different modes and the impact that choice of mode or modes might have on the conclusions from a study.

Psychological theories of survey response will be considered in depth in *Chapter 2*. However, survey non-response and increasing concerns about maintaining adequate levels of response have led researchers to seek to categorise different forms of non-response. For example, Groves and Couper[4] distinguish non-response due to non-contact, refusal to co-operate and inability to participate. The use of incentives to maintain response has, in turn, fostered theoretical development about how such inducements work, which, for example, have focused upon economic theories of incentives through to models describing a broader consideration of social exchange. Comprehensive theories of survey involvement have also been introduced and tested empirically.[5]

More recently, a paradigm shift has been described within survey methodology from a statistical model focused upon the consequences of surveying error to social scientific models exploring the causes of error.[6] Attempts to develop such theories of (1) survey error, (2) decisions to participate and (3) response construction have been brought under the general banner of the Cognitive Aspects of Survey Methodology (CASM) movement. Understanding and reducing measurement error, rather than sampling error, is at the forefront of this endeavour. An impetus for recent theoretical developments is very much provided by technological innovation and diversity, and a requirement to understand the relative impact of different data collection modes upon survey response.

Several information-processing models describing how respondents answer questions have been proposed, which share a common core of four basic stages: (1) comprehension of the question; (2) retrieval of information from autobiographical memory; (3) use of heuristic and decision processes to estimate an answer; and (4) response formulation.[7] These models describe mostly sequential processing. A good example of a sequential information processing model is provided by Tourangeau *et al.*[8] For each stage, there are associated processes identified, which a respondent may or may not use when answering a question. Each stage and each process may be a source of response error.

As indicated above, there has been a substantial expansion in the modes of data elicitation and collection available to survey researchers over the last 30 years. In 1996, Tourangeau and Smith[9] identified six methods that may be used.[9] A quick look at the literature since then will show that this expansion has continued with measures utilised that include personal digital assistants (PDAs) and websites. Subsequently, Tourangeau *et al.*[8] delineated 13 different modes of survey data collection (including remote data collection methods such as telephone, mail, e-mail and the internet), which they considered differed in terms of five characteristics: (1) how respondents were contacted; (2) the presentational medium (e.g. paper or electronic); (3) method of administration (via interviewer or self-administered); (4) sensory input channel used; and (5) response mode.[8]

Variations even within the same mode of data collection further complicate comparison. For example, Honaker[10] describes computer-administered versions of the Minnesota Multiphasic Personality Inventory (MMPI), which differ in terms of type of computer being used, different computer–user interfaces with inconsistent item presentation and response formats. Therefore, different computerised versions of a test cannot be easily generalised to other versions. Other variables that could mediate the effect of different modes of data collection have also been considered, including the overall pace of the interview, the order of survey item processing and the role of different mental models used by respondents. Although the role of different mental models used by respondents, in particular, is rarely assessed, it has been considered a potentially significant mediator of response behaviour.[8]

The challenge for health sciences research

As described above, the first characteristic underlying the different modes of data collection considered by Tourangeau *et al.*[8] was method of contact. Work assessing the impact of an integrated process of respondent approach, consent and data collection has addressed bias due to selective non-ascertainment (i.e. the exclusion of particular subgroups). This may be clearly identifiable subgroups, in terms of people without telephones or computers (for telephone or internet approaches), or less clearly identifiable subgroups, i.e. those with lower levels of literacy or the elderly (for paper-based approaches). There is also considerable work on improving response rates and the biases induced by certain subgroups being less likely to consent to take part in a survey.

Furthermore, an important question in health services research is the use of data collection methods within prospective studies, where patients have already been recruited via another approach. This could be within a clinic or other health service setting rather than the survey instrument being the method of approach as well as data collection. Edwards *et al.*[3] have recently updated a review of the literature (both health and non-health) to identify randomised trials of methods of improving response rates to postal questionnaires. Another review in health-related research has focused on the completeness of data collection and patterns of missing data, as well as response rates.[2]

Guidance is needed not just about the 'best' method to use and most appropriate theoretical model of response, but also the consequence of combining data collected via different modes. For example, a common multimethod approach is when a second mode of data collection is used when the first has been unsuccessful (e.g. using telephone interview when there has been no response to a postal approach[11]). Criteria for judging equivalence of the two approaches are therefore required. Honaker[10] uses the concepts of *psychometric equivalence* and *experiential equivalence*. The former describes when the two forms produce results with equal mean scores, identical distribution and ranking of scores and agreement in how scores correlate with other variables. The latter deals with how two forms may differ in how they affect the psychometric and non-psychometric components of the response task.

In order to inform health services research, guidance is needed which quantifies the differences between modes of data collection and indicates which factors are associated with the magnitude of this difference. These could be *contextual-based* in terms of where the participant is when the information is completed (e.g. health setting, own home, work), *content based* in terms of questionnaire topic (e.g. attitudes to sexual behaviour) or *population based* (e.g. elderly). The factors identified by Tourangeau *et al.*[8] also need to be tested across a wide range of modes and studies.

Aim

The aim of this project is to identify generalisable features affecting responses to the different modes of data collection relevant to health research from a systematic review of the literature.

Objectives

- To provide an overview of the theoretical models of survey response and how they relate to health research.

- To review all studies comparing two modes of administration for subjective outcomes and assess the impact of mode of administration on response quality.
- To explore the impact of findings for key identified health-related measures.
- To create an accessible resource for health science researchers, which will advise on the impact of the selection of different modes of data collection on response.
- To inform the analysis of multimode studies.

Chapter 2

Theoretical perspectives on data collection mode

Background

Understanding the unique experience of both users and providers of health services requires a broad range of suitably robust qualitative and quantitative methods. Both observational (e.g. epidemiological cohort) and interventional studies [e.g. randomised controlled trials (RCTs)] may collect data in a variety of ways, and often require self-report from study participants. Increasingly in clinical studies, clinical indicators and outcomes will form part of an assessment package in which patient lifestyle choices and behaviour, attitudes and satisfaction with health-care provision are a major focus. Health researchers need both to be reassured and to provide reassurance that the measurement tools available are fit for purpose across a wide range of contexts. This applies not only to the survey instrument itself, but also to the way it is delivered and responded to by the participant.

Options for collecting quantitative self-reported data have expanded substantially over the last 30 years, stimulated by technological advances in telephony and computing. The advent of remote data capture has led to the possibility of clinical trials being conducted over the internet.[12,13] Concerns about survey non-response rates have also led researchers to innovate – resulting in greater diversity in data collection.[14] Consequently, otherwise comparable studies may use different methods of data collection. Similarly, a single study using a sequential mixed-mode design may involve, for example, baseline data collection by self-completion questionnaire and follow-up by telephone interview. This has led to questions about the comparability of data collected by the different methods.[15]

In this chapter we apply a conceptual framework to examine the differences generated by the use of different modes of data collection. Although there is considerable evidence about the effect of different data collection modes upon response rates, the chapter addresses the processes that may ultimately impact upon response quality.[16–19] The framework draws upon an existing cognitive model of survey response by Tourangeau *et al.*,[8] which addresses how the impact of different data collection modes may be mediated by key variables. Furthermore, the chapter extends the focus of the model to highlight specific psychological response processes that may follow initial appraisal of survey stimulus. Although much of the empirical evidence for mode effects has been generated by research in other sectors, the relevance for health research will be explored. In doing so, other mediators of response will be highlighted.

It is important to clarify what lies outside the scope of the current review. Although mode of data collection can impact upon response *rate* as well as response *content*, that is not the focus of this report. Similarly, approaches that integrate modes of data collection within a study or synthesise data collected by varying modes across studies are addressed only in passing. Although these are important issues for health researchers, this review concentrates on how the mode of data collection affects the nature of the response provided by respondents, with a particular emphasis on research within the health sciences.

Variance attributable to measurement method rather than the intended construct being measured has been well recognised in the psychological literature and includes biases such as social desirability and acquiescence bias.[20] This narrative review has been developed alongside the systematic literature review of mode effects in self-reported subjective outcomes presented in the subsequent chapters.[21] The chapter highlights for researchers how different methods of collecting self-reported health data may introduce bias and how features of the context of data collection in a health setting such as patient role may modify such effects.

Modes and mode features

What are modes?

Early options for survey data collection were either face-to-face interview, mail or telephone. Evolution within each of these three modes led to developments such as computer-assisted personal interview (CAPI), web-delivered surveys and interactive voice response (IVR). Web-based and wireless technologies, such as mobile- and PDA-based telephony, have further stimulated the development of data collection methods and offer greater efficiency than traditional data collection methods, such paper-based face to face interviews.[22] Within and across each mode a range of options are now available and are likely to continue expanding.

A recent example of technologically enabled mode development is computerised adaptive testing (CAT). Approaches such as item response theory (IRT) modelling allow for survey respondents to receive differing sets of calibrated question items when measuring a common underlying construct [such as health-related quality of life (HRQoL)].[23] Combined with technological advances, this allows for efficient individualised patient surveys through the use of computerised adaptive testing.[24] In clinical populations, CAT may reduce response burden, increase sensitivity to clinically important changes and provide greater precision (reducing sample size requirements).[23] Although IRT-driven CAT may be less beneficial where symptoms are being assessed by single survey items, more general computer-aided testing that mimics the normal clinical interview may be successfully used in combination with IRT-based CAT.[25]

What are the key features of different data collection modes?

The choice of mode has natural consequences for how questions are worded. Face-to-face interviews, for example, may use longer and more complex items, more adjectival scale descriptors and show cards.[26] In contrast, telephone interviews are more likely to have shorter scales, use only end-point descriptors and are less able to use visual prompts, such as show cards. However, even when consistent question wording is maintained across modes there will still be variation in how the survey approach is appraised psychologically by respondents.

The inherent complexity of any one data collection approach (e.g. the individual characteristics of a single face-to-face interview paper-based survey) and increasing technological innovation means that trying to categorise all approaches as one or other mode may be too simplistic. Attention has therefore been focused upon survey design features that might influence response. Two recent models by Groves *et al.*[18] and Tourangeau *et al.*[8] illustrate this. Tourangeau identified five features: (1) how respondents were contacted (e.g. by post, in person); (2) the presentational medium (e.g. paper or electronic); (3) method of administration (interviewer- or self-administered); (4) sensory input channel (e.g. visual or aural); and (5) response mode (e.g. handwritten, keyboard, telephone).[27] Groves *et al.*[18] also distinguished five features: degree of interviewer involvement, level of interaction with respondent, degree of privacy, channels of communication (i.e. sensory modalities) and degree of technology.[28] Although both models cover similar ground, Groves *et al.*[18] place a greater emphasis upon the nature of the relationship between the respondent and the interviewer. Both models attempt to isolate the active ingredients of survey mode. However, Groves *et al.*[18] note that in practice differing combinations of features make generalisation difficult – reflected in their emphasis upon each individual feature being represented as a continuum. Although research on data collection methods has traditionally referred to as 'mode', given the complexity highlighted above, where appropriate we use the term 'mode feature' in this chapter.

How mode features influence response quality

Common to several information-processing models of how respondents answer survey questions there are four basic stages: (1) comprehension of the question; (2) retrieval of information from autobiographical memory; (3) use of heuristic and decision processes to estimate an answer; and (4) response formulation.[7] At each stage, a respondent may use certain processes when answering a question, which may result in a response error.

Of the features that might vary across data collection method, Tourangeau *et al.*[8] proposed four features that may be particularly influential in affecting response: (1) whether a survey schedule is self-administered or interviewer administered; (2) the use of a telephone; (3) computerisation; and (4) whether survey items are read by (or to) the respondent.[8] Although this chapter focuses on differences between these broad mode features, there may still be considerable heterogeneity within each. For example, computerisation in the form of an individual web-delivered survey may apparently provide a standardised stimulus (i.e. overall package of features) to the respondent, but different hardware and software configurations for each user may violate this assumption.[22]

Tourangeau *et al.*[8] considered three variables to *mediate* the impact of mode feature: degree of impersonality, the sense of legitimacy engendered by the survey approach and the level of cognitive burden imposed. Both impersonality and legitimacy represent the respondent's perceptions of the survey approach and instrument. Cognitive burden, impersonality and legitimacy are a function of the interaction between the data collection method and the individual respondent (and subject to individual variation). Nevertheless, the level of cognitive burden experienced by individuals is less dependent upon the respondent's psychological appraisal of the survey task than perceptions of either impersonality or legitimacy.

The relationships among these mode features, mediating variables and three response quality indicators (rate of missing values, reliability and accuracy) are shown in *Figure 1* and have been previously described by Tourangeau *et al.*[8] In this chapter, we further distinguish between psychological appraisals and psychological responses. Psychological appraisals entail the initial processing of salient features by individual respondents and incorporate the mediators described by Tourangeau *et al.* Two additional appraisal processes are included (*leverage–saliency* and *social exchange*) and are described below. Initial appraisal then moves onto psychological response processes. In this amended model, these processes include *optimising/satisficing*, *impression management* and *acquiescence*.[29] Each of these processes is described below and together they represent differing theoretical explanations for an individual's response. The extent to which they are distinct or related processes is also examined.

Other features may also modify response and are added to the chapter framework. They include features of the 'respondent' (the information provider) and 'construct' (what is being measured). These features are not directly related to the method of data collection. Some of these features are implied by the mediators described by Tourangeau *et al.*[8] (e.g. the sensitivity of the construct is implicit to the importance of 'impersonality'). Nevertheless, we consider it important to separate out these features in this framework. Examples of both sets of features are provided, but are intended to be indicative rather than exhaustive listings. Finally, although there may be no

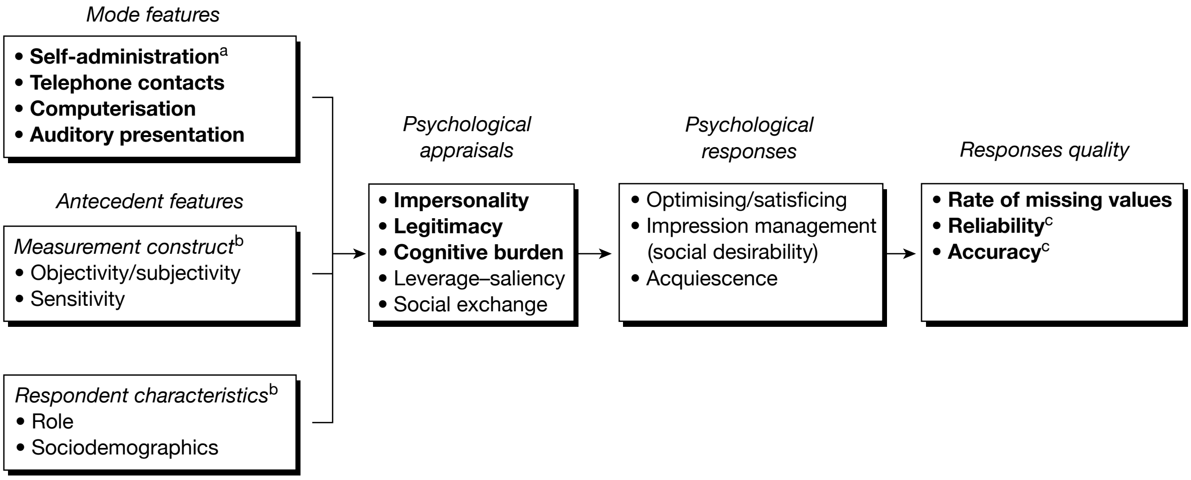


FIGURE 1 Mode features and other antecedent features influencing response quality. (a) Components from Tourangeau's model of impact of data collection mode shown in bold text (Tourangeau *et al*.[8]). (b) Examples from both groups of features are presented. (c) Impact upon level of reporting, for example, rates of smoking, drinking.

unique feature to distinguish between data collection in health and other research contexts, we have used, where we can, examples of particular relevance to health.

How are data collection stimuli appraised by respondents?

Impersonality

The need for approval may restrict disclosure of certain information. Static or dynamic cues (derived from an interviewer's physical appearance or behaviour) provide a social context that may affect interaction.[30] Self-administration provides privacy during data collection. Thus, Jones and Forrest[31] found greater rates of reported abortion among women using self-administration methods than in personal interview. People may experience a greater degree of privacy when interacting with a computer and feel that computer-administered assessments are more anonymous.[32]

The greater expected privacy for methods such as audio computer-assisted self-interview (ACASI) has been associated with increased reporting of sensitive and stigmatising behaviours.[33] It is therefore possible that humanising a computerised data collection interface (e.g. the use of visual images of researchers within computerised forms) could increase misreporting.[34] For example, Sproull *et al.*[35] found higher social desirability scores among respondents to a human-like computer interface compared with a text-based interface. However, others have found little support for this effect in social surveys.[34] Certain data collection methods may be introduced specifically to address privacy concerns – for example, IVR and telephone ACASI. However, there is also evidence that computers may reduce feelings of privacy.[36] The need for privacy will vary with the sensitivity of the survey topic. Although Smith[37] found the impact of the presence of others in response to the US General Social Survey to be mostly negligible, some significant effects were found. For example, respondents rated their health less positively when reporting in the presence of others than when lone respondents.

Legitimacy

Some methods restrict opportunities for establishing researcher credentials, for example when there is no interviewer physically present. A respondent's perception of survey legitimacy could also be enhanced, albeit unintentionally, by the use of computers. Although this may be only a transient phenomenon, as computers become more familiar as data collection tools, other technological advances may produce similar effects (e.g. PDAs).

Cognitive burden

Burden may be influenced by self-administration, level of computerisation and the channel of presentation. Survey design that broadly accommodates the natural processes of responding to questions across these features is likely to be less prone to error.

Leverage–saliency theory

This general model of *survey participation* was proposed by Groves *et al.*[5] and evaluates the balance of various attributes contributing to a decision to participate in a survey. Each attribute (e.g. a financial incentive) varies in importance (leverage) and momentary salience to an individual. Both leverage and salience may vary with the method of data collection and interact with other attributes of the survey (e.g. item sensitivity). Thus, face-to-face interviewers may be able to convey greater salience to responders through tailoring their initial encounter. This common thread of the presence of an interviewer may enhance the perceived importance of the survey to a respondent, which, first, may increase their likelihood of participating (response rate) and, second, enhance perceived legitimacy (response quality). The former effect – 'participation decisions alone' – is not examined further in this review. It is possible that the latter effect of mode feature on response quality may be particularly important in clinical studies if data are being collected by face-to-face interview with a research nurse, for example, rather than by a postal questionnaire.

Social exchange theory

This theory views the probability of an action being completed as dependent upon an individual's perception of the rewards gained and the costs incurred in complying, and his or her trust in the researcher. Dillman[38] applied the theory to explaining response to survey requests – mostly in terms of response rate, rather than quality. However, he noted how switching between different modes within a single survey may allow greater opportunities for communicating greater rewards, lowering costs and increasing trust. This focus upon rewards may become increasingly important as response rates in general become more difficult to maintain. Furthermore, the use of a sequential mixed-mode design for non-respondent follow-up within a survey may enhance perceptions of the importance of the research itself by virtue of the researcher's continued effort.

Unlike the first three appraisal processes described above, both leverage–saliency and social exchange address broader participation decisions. Features of different data collection modes may affect such decision-making, for example through perceived legitimacy. Other features in the framework considered to modify response may also influence participation decisions according to these theories (e.g. the sensitivity of the construct being measured).

Explaining mode feature effects: psychological responses following appraisal

Initial appraisal of survey stimulus will result in a response process, which further mediates response quality. Several explanatory psychological theories have been proposed. We focus upon three general theories of response formulation (optimising/satisficing, social desirability and acquiescence).

'Taking the easy way out' – optimising and satisficing

Krosnick[29,39] described 'optimising' and 'satisficing' as two ends of a continuum of thoroughness of the response process. Full engagement in survey response represents the ideal response strategy (optimising), in contrast to incomplete engagement (satisficing). The theory acknowledges the cognitive complexity of survey responding. A respondent may proceed through each cognitive step less diligently when providing a survey response or may omit *information retrieval* and *judgement* completely (examples of weak and strong satisficing, respectively). In either situation, respondents may use a variety of decision heuristics when responding. Three factors are considered to influence the likelihood of satisficing: respondent ability, respondent motivation and task difficulty.[29,40] Krosnick[39] defines respondent ability (or cognitive sophistication) as the ability to retrieve information from memory and integrate it into verbally expressed judgements. Optimising occurs when respondents have sufficient cognitive sophistication to process the request, when they are sufficiently motivated and when the task requirements are minimal.[42]

Mode feature effects may influence optimising through differences in non-verbal communication, interview pace (speed) and multitasking. First, the enthusiastic non-verbal behaviour of an interviewer may stimulate and maintain respondent motivation. Experienced interviewers react to non-verbal cues (e.g. expressions relating to lack of interest) and respond appropriately. Such advantages are lost in a telephone interview with interviewers relying on changes in verbal tones to judge respondent engagement. Although the role of an interviewer to enhance the legitimacy of the survey request was highlighted in Tourangeau *et al.*'s[8] framework, this additional motivation and support function was not clarified. Second, interview pace may differ between telephone and face-to-face contact, in part because silent pauses are less comfortable on the telephone. A faster pace by the interviewer may increase the task difficulty (cognitive burden) and encourage respondents to take less effort when formulating their response. Pace can vary between self- and interviewer-administered methods. A postal questionnaire may be completed at respondents' own pace, allowing them greater understanding of survey questions compared with interviewer-driven methods. Tourangeau *et al.*[8] omitted pace as a mediating variable from their model of mode effects because they considered that insufficient evidence had accrued to support its role. Interview pace has been suggested as an explanation for observed results, but the effects of pace have not necessarily been tested independently from other mode features (e.g. see Kelly *et al.*[43]). Nevertheless, it is discussed here because of its hypothesised effect.[29] Finally, distraction due to respondent multitasking may be more likely in telephone interviews than in face-to-face interviews (e.g. telephone respondents continuing to interact with family members or conduct household tasks while on the telephone). Such distraction increases task difficulty and thus may promote satisficing.[29]

Optimising/satisficing has been used to explain a variety of survey phenomena, for example response order effects (where changes in response distributions result from changes in the presentational order of response options).[44] Visual presentation of survey questions with categorical response options may allow greater time for processing initial options leading to primacy effects in those inclined to satisfice. Weak satisficing may also result from the termination of evaluative processing (of a list of response options) when a reasonable response option has been encountered. This may occur for response to items with adjectival response scales and also for ranking tasks.[29] In contrast, aural presentation of items may cause respondents to devote more effort to processing later response options (which remain in short-term memory after an interviewer pauses), leading to recency effects in satisficing respondents.[41] Telephone interviews can increase satisficing (and social desirability response bias) compared with face-to-face interviews.[42] An example of a theoretically driven experimental study that has applied this parsimonious model to studying mode feature effects is provided by Jäckle *et al.*[45] In the setting of an interviewer-delivered social survey, they evaluated the impact of question stimulus (with or without show cards) and the physical presence or absence of interviewer (face to face or telephone). In this instance, detected mode feature effects were attributable not to satisficing, but to social desirability bias instead.

Social desirability

The tendency for individuals to present themselves in a socially desirable manner in the face of sensitive questions has long been inferred from discrepancies between behavioural self-report and documentary evidence. Response effects due to self-presentation are more likely when respondents' behaviour or attitudes differ from their perception of what is socially desirable.[46]

This may result in over-reporting of some behaviours and under-reporting of others. Behavioural topics considered to induce over-reporting include being a good citizen and being well informed and cultured.[47] Under-reporting may occur with certain illnesses (e.g. cancer and mental ill-health), illegal and non-normative behaviours and financial status. An important distinction has been made between intentional impression management (a conscious attempt to deceive) and unintentional self-deception (where the respondent is unaware of his or her behaviour).[48] The former has been found to vary according to whether responses were public or anonymous, whereas the latter was invariant across conditions.

Most existing data syntheses of mode effects have related to social desirability bias (*Table 1*). Sudman and Bradburn[46] indicated the importance of the method of administration upon socially desirable responding. They found a large difference between surveys either telephone- or self-administered compared with face-to-face interviews. Differences in social desirability between modes have been the subject of subsequent meta-analyses by de Leeuw,[49] Richman *et al.*[50] and Dwight and Feigelson.[51] De Leeuw[49] analysed 52 studies, conducted between 1947 and 1990, comparing telephone interviews, face-to-face interviews and postal questionnaires. There was no overall difference in socially desirable responding between face-to-face and telephone surveys among 14 comparisons. There was, however, more bias in telephone interviews in the nine studies published before 1980, but no difference in the later studies. There was less socially desirable responding in postal surveys than in both face-to-face surveys (13 comparisons, mean $r=0.09$) and telephone surveys (five comparisons, mean $r=0.06$). The presence of an interviewer (telephone or face to face), therefore, appears to determine socially desirable responding. The review included both subjective and objective outcomes, and health issues were the most prominent topic covered.

The meta-analysis of Richman *et al.*[50] compared computer-administered questionnaires, paper-and-pencil questionnaires and face-to-face interviews in 61 studies. Controlling for moderating factors, there was less social desirability bias in computer administration than in paper-and-pencil administration [effect size (ES) for difference of 0.39]. This advantage over paper-and-pencil methods was greater in studies conducted before 1975 (ES = 0.74), when responses were provided when alone (ES = 0.82) and when backtracking (i.e. ability to move back to earlier section of questionnaire) was available (ES = 0.65). However, when social desirability was inferred from other measures (rather than measured directly) there was more bias using computer administration controlling for moderators (ES = 0.46). Compared with face-to-face interviews, computer administration was associated with less bias overall (ES = 0.19). However, the opposite was true when the construct assessed was personality (ES = 0.73) and in more recently published studies (ES = 0.79).

Dwight and Feigelson[51] compared impression management/self-deceptive enhancement in computer-administered measures and either paper-and-pencil or face-to-face measures. Less impression management bias was found for computer administration than for non-computer formats, but the difference was small (ES = –0.08). Individual study ESs reduced significantly over time, indicating a diminishing impact of computerisation. Dwight and Feigelson[51] pointed to the recent positive ESs, which they felt was consistent with a 'Big Brother syndrome' – respondents fear monitoring and controlling by computers.[52] There was no observed difference between data collection method on scores of self-deceptive enhancement.

It is worth commenting upon the methodological quality of these reviews.[53] None provided an explicit search strategy, although all, apart from Sudman and Bradburn,[47] described keywords. Dwight and Feigelson's[52] search was based upon an initial citation search, whereas only Richman *et al.*'s[51] review provided explicit eligibility criteria for included studies. Sudman and Bradburn[46] developed a comprehensive coding scheme that was later extended in de Leeuw's review.[49]

TABLE 1 Reviews of mode effects in socially desirable responding

Review details	Modes compared	No. of comparisons	Primary result	Evidence of effect moderators
Sudman and Bradburn	Face to face, self-administration			
Years: not reported[a]	1. Strong possibility of SD answer		RE: face to face = 0.19, self-administration = 0.32	
Effect estimate: relative RE	2. Some possibility of SD answer		RE: face to face = 0.11, self-administration = 0.22	
Studies: $n = 305$[a]	3. Little/no possibility of SD answer		RE: face to face = 0.15, self-administration = 0.19	
Commentary: The effect measure for attitudinal variables compares any one mode with the weighted mean of all responses (not a direct mode vs mode comparison). Differences in size of RE indicate that one mode has more/less bias than another, but not how much. Individual sample size not accounted for in analysis and may have created spurious results				
De Leeuw *Years*: 1947–1990	1. Telephone vs face to face	$n = 14$	No overall difference (mean = –0.01)	*Year of publication*: '< 1980' (mean = –0.03; less bias by face to face), 'after 1980' (mean = 0.00)
Effect estimate: mean weighted product moment correlation	2. Mail q vs face to face	$n = 13$	Less bias by mail (mean = +0.09)	
Studies: $n = 52$[a]	3. Mail q vs telephone	$n = 5$	Less bias by mail (mean = +0.06)	
Commentary: The square of the correlation indicates proportion of variance explained by mode. The directional coefficient indicates which mode is best (less biased). 'Social desirability' assessed by authors of original papers, not review paper				
Richman *et al.* *Years*: 1967–1997	1. Computer vs PAPQ (studies – BS: $n = 30$; WS: $n = 15$)	$n = 581$	No overall difference (ES = 0.05)	
Effect estimate: ES *Studies*: $n = 61$	a. Direct measure of bias		Less bias by computer (ES = –0.39)	*Year of publication*: 'early:1975' (ES = –0.74); 'recent: 1996' (ES = –0.08) *Alone*: 'alone' (ES = –0.82); 'not alone' (ES = –0.25) *Backtracking*: available (ES = –0.65); not available (ES = –0.24)
			No difference in effect between IM and SDE bias	
	b. Inferred measure of bias		Less bias by PAPQ (ES = 0.46)	*Anonymity*: 'anonymous' (ES = 0.25); 'identified' (ES = 0.62) *Alone*: 'alone' (0.12); 'not alone' (0.65) *Backtracking*: available (ES = 0.16); not available (ES = 0.87)
	2. Computer vs face-to-face (Studies – BS: $n = 11$; WS: $n = 17$)	$n = 92$	Less bias by computer (ES = –0.19)	*Measure*: personality (ES = 0.73); other (ES –0.51) *Year of publication*: 'early: 1975' (ES = 0.79); recent: 1996 (ES: –1.03)
Dwight and Feigelson *Years*: 1969–1997	1. Computer vs paper and pencil or face to face (studies – BS: $n = 33$; WS: $n = 30$)	IM: $n = 45$ SDE: $n = 32$	Less IM bias by computer (ES = –0.08) No difference in SD bias	Overall ESs for IM bias reduce over time ($r = 0.44$)
Effect estimate: ES *Studies*: $n = 30$	2. Computer vs paper and pencil	IM: $n = 39$ SDE: $n = 6$	Less IM bias by computer (ES = –0.08) No difference in SDE bias	
	3. Computer vs face to face	IM: $n = 25$ SDE: $n = 7$	No difference in SDE bias No difference in SDE bias	

BS, between subjects; IM, impression management; Mail q, mail questionnaire; PAPQ, paper-and-pencil questionnaire; RE, response effect; SD, social desirability; SDE, self-deception enhancement; WS, within subjects.

[a] Includes studies not contributing to social desirability analysis.

However, coding performance (inter-rater reliability) was reported only by de Leeuw[49] and by Richman *et al.*[50] Difficulties in coding variables with their frameworks was noted by Sudman and Bradburn[46] and by Richman *et al.*,[50] but is probably a ubiquitous problem. The intended coverage of the reviews varied where stated, but is probably generally reflected in the total number of included studies. The Richman *et al.*[50] review is notable for its attempt to test explicit a priori hypotheses, its operational definition of 'sensitivity' and its focus upon features rather than overarching modes. These reviews provide support for the importance of self-administration and consequently impersonality. Richman *et al.*[50] concluded that there was no overall difference between computer and paper-and-pencil scales. This is consistent with Tourangeau *et al.*'s[8] model, which directly links computerisation to legitimacy and cognitive burden but not to impersonality. From the first two reviews it is clear that other factors may significantly modify the relationship between mode and social desirability bias. For example, Whitener and Klein[54] found a significant interaction between social environment (individual vs group) and mode of administration (computer:unrestricted scanning vs computer:restricted scanning vs paper-and-pencil).

Acquiescence

Asking respondents to agree or disagree with attitudinal statements may be associated with acquiescence – respondents agreeing with items regardless of there content.[55] Acquiescence may result from respondents taking shortcuts in the response process and paying only superficial attention to interview cues.[18] Knowles and Condon[56] categorise meta-theoretical approaches to acquiescence as addressing either motivational or cognitive aspects of the response process.[3] Krosnick[39] suggested that acquiescence may be explained by the notion of satisficing due to either cognitive or motivational factors. Thus, the role of mode features in varying impersonality and cognitive burden as described above would seem equally applicable here.

There is mixed evidence for a mode feature effect for acquiescence. De Leeuw[49] reported no difference in acquiescence between postal, face-to-face and telephone interviews.[49] However, Jordan *et al.*[57] found greater acquiescence bias in telephone interviews than in face-to-face interviews. Holbrook *et al.*[42] also found greater acquiescence among telephone respondents than among face-to-face respondents in two separate surveys.

What are the consequences of mode feature effects for response quality?

Several mode feature effects on response quality are listed in *Figure 1* and include number of *missing data*.[9] Computerisation and using an interviewer will decrease the number of missing data due to unintentional skipping. Intentional skipping may also occur and be affected by both the impersonality afforded the respondent and the legitimacy of the survey approach. The model of Tourangeau *et al.*[8] describes how the *reliability* of self-reported data may be affected by the cognitive burden placed upon the respondent.[8] De Leeuw[49] provides a good illustration of how the internal consistency (psychometric reliability) of summary scales may be varied by mode features through (1) differences in interview pace and (2) the opportunity for respondents to relate their responses to scale items to each other. The reliability of both multiple- and single-item measures across surveys (and across waves of data collection) may also be affected by any mode feature effects resulting from the psychological appraisal and response processes described above.

Tourangeau *et al.*[8] highlight how *accuracy* (validity) of the data may be affected by impersonality and legitimacy. Both unreliable and inaccurate reporting will be represented by variations in the *level* of an attribute being reported. For example, a consequence of socially desirable responding will be under- or over-reporting of attitudes and behaviour. This may vary depending upon the degree of impersonality and perceived legitimacy.

Additional antecedent features

Two further sets of variables are considered in the framework presented in *Figure 1*: 'measurement construct' and 'respondent characteristics'. Both represent antecedent features that may further interact with the response process described. For the purposes of this chapter they will be described particularly in relation to health research.

Measurement construct

Objective/subjective constructs

Constructs being measured will vary according to whether they are subjective or objectively verifiable. HRQoL and health status are increasingly assessed using standardised self-report measures [increasingly referred to as patient-reported outcome measures (PROMs) in the health domain]. Although the construct being assessed by such measures may in some cases be externally verified (e.g. observation of physical function), for other constructs (e.g. pain) this may not be possible. Furthermore, the subjective perspective of the individual may be an intrinsic component of the construct being measured.[58,59] Cote and Buckley[60] reviewed 64 construct validation studies from a range of disciplines (marketing, psychology/sociology, other business, education) and found that 40% of observed variance in attitudes (subjective variable) was due to method (i.e. the influence of measurement instrument) compared with 30% being due to the trait itself. For more objective constructs, variance due to method was lower indicating the particular challenge for assessing subjective constructs.

Sensitivity

Certain clinical topics are more likely to induce social desirability response bias, potentially accentuating mode feature effects. Such topics include sensitive clinical conditions (e.g. human immunodeficiency virus status) and health-related behaviours (e.g. smoking). An illustrative example is provide by Ghanem *et al.*[61] who found more frequent self-reports of sensitive sexual behaviours (e.g. number of sexual partners in preceding month) using ACASI than with face-to-face interview among attendees of a public sexually transmitted diseases clinic.

Respondent characteristics

Respondent role

In much of the research contributing to the meta-analyses of mode effects on social desirability, the outcome of the assessment was not personally important for study subjects (e.g. participants being undergraduate students).[50] Further methodological research in applied rather than laboratory settings will help determine whether or not mode feature effects are generalisable to wider populations of respondents. It is possible that the motivations of patients (e.g. perceived personal gain and perceived benefits) will reflect their clinical circumstances, as well as other personality characteristics.[62–64] It is therefore worth investigating whether or not self-perceived clinical need, for example, may be a more potent driver of biased responding than social desirability, and whether or not this modifies mode feature effects.

In a review of satisfaction with health care, the location of data collection was found to moderate the level of satisfaction reported, with on-site surveys generating less critical responses.[19] Crow *et al.*[19] noted how the likelihood of providing socially desirable responses was commonly linked by authors to the degree of impersonality afforded when collecting data either on- or off-site.

Another role consideration involves the relationship between respondent and researcher. The relationship between patient and health-care professional may be more influential than that between social survey respondent and researcher. A survey request may be viewed as particularly legitimate in the former case and less so in the latter.[63] Response bias due to satisficing may be less of a problem in such clinical populations than in non-clinical populations. Systematic evaluation of the consequence of respondent role in modifying mode feature effects warrants further research.

Respondent sociodemographics

There is some indication of differential mode feature effects across demographic characteristics. For example, Hewitt[65] reports variation in sexual activity reporting between modes [audio-computer-assisted self-administered interview (CASI) and personal interview] by age, ethnicity, educational attainment and income. The epidemiology of different clinical conditions will be reflected by patient populations that have certain characteristics, for example being older. This may have consequences for cognitive burden or perceptions of legitimacy in particular health studies.

Particular issues in health research

In considering modes and mode feature effect, we will focus on three issues that may be of particular relevance to those collecting data in a health context: antecedent features, constraints in choice of mode and the use to which the data are being put.

Particular antecedent features

Certain antecedent conditions and aspects of the construct being measured may be particularly relevant in health-related studies. Consider the example of QoL assessment in clinical trials of palliative care patients from the perspective of response optimising. Motivation to respond may be high, but may be compromised by an advanced state of illness. Using a skilled interviewer may increase the likelihood of optimising over an approach offering no such opportunity to motivate and assist the patient. Physical ability to respond (e.g. verbally or via a keyboard) may be substantially impaired. This may affect response completeness, but if the overall response burden (including cognitive burden) is increased it may also lead to satisficing. In practice, choice of data collection method will be driven as much by ethical considerations about what is acceptable for vulnerable respondents.

Are there features of self-reported data collection in health that are particularly different from other settings of relevance to mode feature effects?

Surveys will be applied in health research in a wide variety of ways, and some will be indistinguishable in method from some social surveys (e.g. epidemiological sample surveys). Some contexts for data collection in health research may be very different from elsewhere. Data collection in RCTs of therapeutic interventions may often include PROMs to assess differences in outcome. How antecedent features in the trial – in particular those associated with respondent role – may influence psychological appraisal and response is hypothesised in *Table 2*. These antecedent characteristics may potentially either promote or reduce the adverse impact of mode feature effects. The extent to which these effects may be present will need further research, and, at least, would require consideration in trial design.

Particular constraints on choice of mode

As in social surveys, mode feature effects will be one of several design considerations when collecting health survey data. Surveying patients introduces ethical and logistical considerations, which, in turn, may determine or limit the choice of data collection method. Quality criteria such as appropriateness and acceptability may be important design drivers.[66] For example, Dale and Hagen[67] reviewed nine studies comparing PDAs with pen-and-paper methods and found higher levels of compliance and patient preference with PDAs. Electronic forms of data collection may offer advantages in terms of speed of completion, decreasing patient burden and enhancing acceptability.[68,69] The appropriateness of different data collection modes may vary by patient

TABLE 2 How mode and antecedent features may influence response: the example of respondent role in a clinical trial

Antecedent features in trial	Appraisal and response: some research hypotheses
Respondent role: Participants approached for participation by their professional carer	*Legitimacy*: An established patient–carer relationship with high levels of regard for the researcher may enhance legitimacy of the survey request sufficiently to modify mode feature effects and therefore reduce satisficing
Respondent role: Participants are consented through a formally documented process	*Legitimacy*: The formality and detail of the consenting process may enhance the legitimacy of the survey request sufficiently to modify mode feature effects and therefore reduce satisficing
Respondent role: Participants provide self-reported data at the site of delivery for their health care	*Impersonality*: On-site data collection may increase the need for confidential and anonymous reporting sufficiently to promote adverse effects of mode feature effects and introduce social desirability bias
Respondent role/sensitivity: Participants are patients with an ongoing clinical need	*Cognitive burden*: The health status of respondent may increase the overall cognitive burden to modify mode feature effects and increase satisficing. Burden and, therefore, effects may vary with disease and treatment progression *Impersonality*: The nature of the condition may increase the need for confidential and anonymous reporting sufficiently to promote adverse mode feature effects and introduce social desirability bias
Respondent role: Participants are patients in receipt of therapeutic intervention	*Legitimacy/leverage–saliency*: The requirement for treatment and the opportunity for novel therapy enhance legitimacy and the perceived importance/salience of the research. This may minimise adverse mode feature effects to reduce satisficing

group – for example, with impaired response ability due to sensory loss.[70] Health researchers need to balance a consideration of mode feature effects with other possible mode constraints when making decisions about data collection methods.

Particular uses of data

Evaluating mode feature effects will be particularly important as survey instruments start to play a bigger role in the provision of clinical care, rather than solely in research. PROMs are increasingly being applied and evaluated in routine clinical practice.[71–73] Benefits have been found in improving process of care, but there is less consistent evidence for impact on health status.[71,74–76]

Perceived benefits of using such patient-reported outcomes include assessing the impact on patients of health-care interventions, guiding resource allocation and enhancing clinical governance.[72] Computerised data collection may be especially important if results are to inform actual consultations, but would require suitably supported technology to permit this.[77,78] With only mixed evidence of clinical benefit, Guyatt *et al.*[76] highlight computerised-based methods of collecting subjective data in clinical practice as a lower-cost approach.

In this clinical service context, psychological responses such as social desirability bias may vary according to whether patient data are being collected to inform treatment decision-making or clinical audit. Method of data collection may similarly play a role in varying the quality of response provided. However, routinely using subjective outcome measures in clinical practice will require a clear theoretical basis for their use and implementation, and may necessitate additional training and support for health professionals and investment in the technology to support its effective implementation, which is, preferably, cost neutral.[79–82] Overall, though, it may be that any biasing effect of mode feature may be less salient in situations where information is being used as part of a consultation to guide management, and may be more so where data are being collected routinely across organisational boundaries as part of clinical audit or governance.

Managing mode feature effects in health

Managing mode feature effects requires identification of their potential impact. This chapter has focused upon response quality as one source of error in data collection. Two other sources of error influenced by mode are 'coverage error' and 'non-response error'.[83] In the former, bias may be introduced if some members of the target population are effectively excluded by features of the chosen mode of data collection. For example, epidemiological surveys using random digit dialling, which exclude people without landline telephones, may result in biased estimates as households shift to wireless-only telephones.[84] Response rates vary by mode of data collection and different population subgroups vary in the likelihood of responding to different modes.[83] For example, Chittleborough *et al.*[85] found differences by education, employment status and occupation among those responding to telephone and face-to-face health surveys in Australia.

Social surveys commonly blend different modes of data collection to reduce cost (e.g. by a graduated approach moving from cheaper to more expensive methods[18]). Mixing modes can also maximise response rates by, for example, allowing respondents a choice about how they respond.

In the long term it may prove possible to correct statistically for mode feature effects if consistent patterns emerge from meta-analyses of empirical studies. Alternatively, approaches to reducing socially desirable responding have targeted both the question threat and confidentiality. An example of the latter is the randomised response technique, which guarantees privacy.[86,87] Another approach is the use of goal priming (i.e. the manipulation and activation of individuals' own goals to subsequently motivate their behaviour), where respondents are influenced subconsciously to respond more honestly.[88]

Evaluating and reporting mode feature effects

As described above, the evaluation of data collection method within individual studies is usually complicated by the package of features representing any one mode. Groves *et al.*[18] described two broad approaches to the evaluation of effects due to mode features. The first and more pragmatic strategy involves assessing a package of features between two or more modes. Such a strategy may not provide a clear explanation for resulting response differences, but may satisfy concerns about whether or not one broad modal approach may be replaced by another. The second approach attempts to determine the features underlying differences found between two modes. This theoretically driven strategy may become increasingly important as data collection methods continue to evolve and increase in complexity.

As global descriptions of data collection method can obscure underlying mode features, comparative studies should describe these features more fully. This would enable data synthesis, providing greater transparency of method and aid replication.[50]

Summary

This chapter has considered how features of data collection mode may impact upon response quality, and key messages are summarised in *Box 1*. It has added to a model proposed by Tourangeau *et al.*[8] by drawing apart psychological appraisal and response processes in mediating the effect of mode features. It has also considered other antecedent features that might influence response quality. Mode feature response effects have been most thoroughly reviewed empirically in relation to social desirability bias. Overall effects have been small, although evidence of significant effect modifiers emphasises the need to evaluate mode features rather than simply overall mode. A consistent finding across the reviews is the important moderating effect of year of publication for comparisons involving both telephone and computers. Therefore, mode feature comparisons are likely to remain important as new technologies emerge for collecting data. Although much of the empirical research underpinning the reviewed model has been generated

BOX 1 Key messages for researchers and for the systematic review

Broad messages for researchers

Choice of data collection mode can introduce measurement error, detrimentally affecting the accuracy and reliability of survey response

Surveys in health service and research possess similar features to surveys in other settings

Features of the clinical setting, the respondent role and the health survey content may emphasise psychological appraisal and psychological responses implicated in mode feature effects

The extent to which these features of health surveys result in consistent mode effects that are different from other survey context requires further evaluation

Evaluation of mode effects should identify and report key features of data collection method, not simply categorise by overall mode

Mode feature effects are primarily important when data collected via different modes are combined for analysis or interpretation. Evidence for consistent mode effects may nevertheless permit routine adjustment to help manage such effects

Implications for the MODE ARTS systematic review

The theory review provides the framework to structure the systematic review analysis

In doing so it emphasises mode features, rather than modes

Other antecedent features identified in the review may also be explored in the analysis, which, in themselves, may not be directly associated with mode feature

Mediators are clarified in the theoretical framework, but in practice clear measures of these are unlikely to be available in the papers obtained in the systematic review. Nevertheless, the theoretical framework provides a firm basis upon which to interpret emergent results

MODE ARTS, Mode of Data Elicitation, Acquisition & Response To Surveys.

within other academic domains, the messages are nonetheless generally applicable to clinical and health research.

Future evidence syntheses may confirm or amend the proposed model, but this requires as a precursor greater attention to theoretically driven data collection about mode features. The current theoretical review framework, therefore, provides the basic analytic structure for the analysis and a basis upon which emergent results may be interpreted (see *Box 1*). In particular, the emphasis upon mode features is a key contributor to this analytic model.

Chapter 3

Review methods

The methods used to evaluate the impact of features of mode of data collection on subjective outcome measures follow that of a systematic review. Owing to the large body of literature relating to survey methodology which is published outside the health research arena, all studies that incorporate a mode comparison will be included, regardless of setting. This leads to a broad search strategy covering a wide range of disciplines. In order to target methodological studies, some innovations in search strategy have been undertaken that separate out the process from the traditional reviews of the effectiveness of interventions.

Identifying the literature

Inclusion/exclusion criteria

The inclusion/exclusion criteria for a study to be included in the review were as follows:

- There is evidence of a comparison between two modes of data collection of either the same question, or set of questions, referring to the same theoretical construct.
- The construct compared is subjective and cannot be externally validated.
- The analysis of the study contains an explicit reference to a comparison.
- Data collection is quantitative, i.e. uses structured questions and answers.

This can include studies in which mode comparisons were made, even if not the main purpose of the study.

Studies were excluded from the review if they involved:

- a comparison between a quantitative measure and one or more qualitative data collection methods/analyses (e.g. unstructured interviews, focus groups)
- a comparator derived from routine clinical records – unless explicit reference to specific self-reported construct is made within those records
- a comparison between the response of two different judges, i.e. comparing a response from an individual with that made by someone other than the responder, for example a clinician providing a diagnosis.

Subjective measures are defined as those in which the perspective of the individual is an intrinsic component of the construct being measured. Comparisons between two different perspectives (even on the same construct) are therefore excluded.

Year of publication

All databases were searched from the earliest point in time until the end of 2004. This was based on the last complete year available to the researchers at the point at which the search was undertaken.

Language and location

No studies were excluded owing to language or country of origin to allow inclusion of as much innovation in design and novel mode application as possible. It is known that the perceived 'gold standard' method of data collection will vary, especially in relation to approaching respondents in their homes.[89] In some cultures, a face-to-face interview is perceived as more acceptable than calling on the telephone.[90] In addition, matters of privacy and over-use of mass marketing schemes have changed the availability of telephone numbers and ability to contact. For example, the use of automated marketing technology in the UK has given rise to 'preference' services offered by telecommunications companies and the Post Office, whereby registered marketing companies cannot gain access to the recipient.

This chapter will document the development, piloting and optimisation of the search strategy, the three-phase selection process and the methods of synthesis for the data extracted.

Databases

Owing to the broad range of disciplines outside the health sector literature which cover survey methodology, a subject-free approach was required to collate evidence from all research. However, databases were chosen based on subject area to allow as comprehensive a selection as possible. A full list of databases used in the review can be found in *Appendix 1* (see *Table 21*).

MEDLINE was used for the initial development and optimisation of the search strategy. It was decided that grey literature and grey databases would not be searched, as the effort required to retrieve such information usually outweighs the gains.[91] Therefore, only journal articles and conference abstracts cited within journal supplements were included in the review process.

Search strategy

Guides for the development and creation of search strategies used in systematic reviews in defined areas have been well described.[92] However, guides do not exist for searching such a diffuse and multidisciplinary literature base. Therefore, the search strategy for the present review was continually optimised using an iterative process. Initially, an extensive development phase was carried out, followed by the main search and retrieval phase.

From previous literature reviews in the area of survey research[2,93] it was shown to be possible to systematically identify a body of literature describing the effects of differences in modes of data collection. However, studies that use only a single mode of data collection are not of interest and, therefore, in order to focus the search strategy, a matrix approach was developed. The matrix was intended to facilitate the search for articles that had two or more modes of data collection. Each column and row of the matrix consisted of a collection of terms relating to a single mode (e.g. postal, survey, mail). Only the off-diagonal terms were considered for inclusion (highlighted cells in *Table 3*). This used Boolean terminology: Group 1 AND (Group 2 OR 3 OR 4 OR 5 OR 6 OR 7

TABLE 3 Illustration of matrix approach to identification

		Mode of data collection							
		1	2	3	4	5	6	7	8
Mode of data collection	1		▨	▨	▨	▨	▨	▨	▨
	2			▨	▨	▨	▨	▨	▨
	3				▨	▨	▨	▨	▨
	4					▨	▨	▨	▨
	5						▨	▨	▨
	6							▨	▨
	7								▨
	8								

OR 8 OR 9 OR 10). For example, this would identify any paper that had terms relating to desktop computer use *and* any one of face to face, paper and pencil, etc.

Initially, 10 different types of data collection mode were identified, which were defined as 'data collection groups'. A list of search terms was generated for each group. From these categorisations, one row and column (representing paper-and-pencil administration) was selected and all abstracts identified (759) were screened and the terms and categorisations tested to see if a more specific search strategy would have identified the same studies. On the basis of this, the data collection groups were revised from 10 to 8 as follows:

1. technology assisted (computer and PDA combined)
2. internet based
3. antonym of technology
4. paper-and-pencil administration (combined with mail)
5. fax administration
6. telephone administration
7. in-person administration
8. unspecified mode.

It became apparent that there was an ordered use of language in all articles, allowing a grammatical framework to be applied to the search terms within the data collection groups. Search terms relating to different modes of data collection could be described as a nominal phrase, consisting of a compound noun and one or more compound adjectives. New modes of data collection have evolved with the creation of new technologies, and, instead of developing new nouns, existing nouns have been modified by the development of compound nouns, qualified by compound adjectives, for example computer-assisted telephone interview (CATI). Search terms generated in the initial searches were allocated to the different data collection groups by linking the compound adjective to the group with which it was most associated. The final search terms for each data collection group are in *Appendix 1*.

Medical subject headings (MeSH) were utilised where available. The specific thesaurus terms used in each database and field codes used to implement the matrix section of the search strategy are detailed in *Appendix 1* (see *Table 21*), concerning health and evidence-based medicine, social sciences, and economics and other, respectively. The use of MeSH can seriously influence the noise in the search strategy (the number, and type of citations retrieved) due to the branching hierarchical classification. For example, when locating articles related to methodological issues the search term 'method' is prolific in the introduction, method, results and discussion (IMRaD)'-constructed abstracts, whereas the more specific term 'methodolog$' searched in the title, abstract and keywords of the article yielded more precise results.

The strategy was implemented in MEDLINE from the beginning of 1966 to the end of 2004, and all articles were subsequently screened for relevance. The screening accompanied an iterative process identifying new research-specific terms. The iterative process generated 24 new nominal phrases that were added to the appropriate groups, and one new group was identified pertaining to the use of video. No clear distinction was developed between online and offline computerised methods, therefore the terms in the internet-based group were merged with the technology-assisted data collection methods. The strategy was then re-implemented to screen for new, previously unidentified articles.

In order to focus the search on studies that were comparisons of modes, rather than just studies that happened to report two modes, the studies identified from the searches above were limited to those that used terms suggestive of a comparison (e.g. comparison, versus, trial, evaluation) and general terms relating to data collection (e.g. administration, survey, assessment). Therefore, only studies that combined all three domains were included for further consideration (*Figure 2*).

Following the successful development of the search strategy within MEDLINE, the same strategy was implemented within all the specified databases, allowing for changes in field codes and thesaurus terms as described in *Appendix 1*.

Citation information and abstracts were downloaded from the selected databases and imported into an EndNote (Thomson Reuters, CA, USA) database. At each stage of the download, the number of articles requested for download and the numbers of articles actually downloaded were checked for consistency. Duplicate citations were removed using the EndNote Version 9 'Find Duplicate' function. Citations were considered duplicates if either:

- the title field exactly matched another citation, *or*
- the author, year, journal, volume, issue and page numbers exactly matched.

Review process

The abstracts (and titles only for some foreign-language papers with no English abstract) were reviewed against the inclusion/exclusion criteria. No assessment was made at this stage as to the subjectivity of the measures presented. Full papers were retrieved for all selected abstracts and then screened again relating to more detailed inclusion criteria relating to the measures used. Papers that were still included were reviewed in full and detailed data extracted (*Figure 3*). The datasheets used for full-paper screening and data extraction are given in *Appendix 2*. The screening and data extraction stages were combined for foreign-language papers.

At each stage, abstracts or papers were reviewed by a single reviewer after a period of training (*Figure 4*). Training for each stage included an assessment of reliability and sensitivity. Training and testing sets of abstract/papers were used. This was repeated for hits from different databases to allow for reassessment with different types of study and abstract layout.

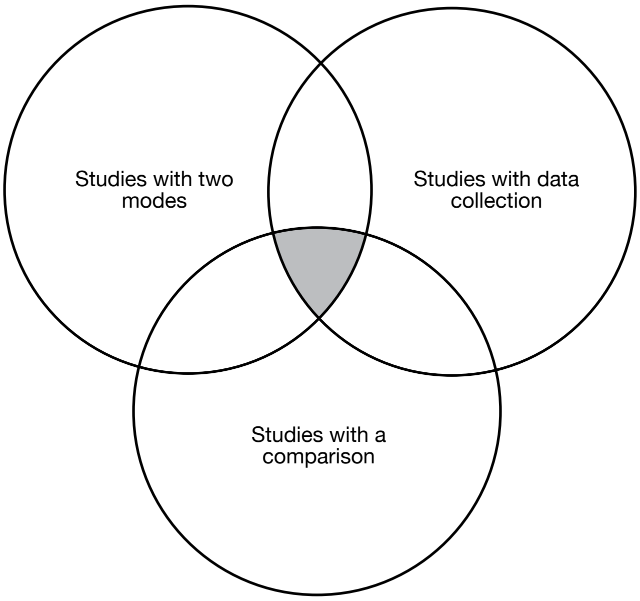


FIGURE 2 Conceptualisation of search strategy.

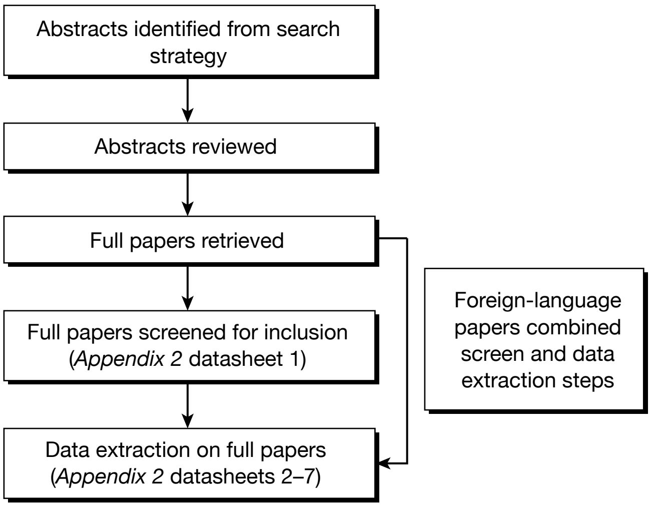

Abstracts identified from search strategy
Abstracts reviewed
Full papers retrieved
Full papers screened for inclusion (*Appendix 2* datasheet 1)
Data extraction on full papers (*Appendix 2* datasheets 2–7)
Foreign-language papers combined screen and data extraction steps

FIGURE 3 Review process.

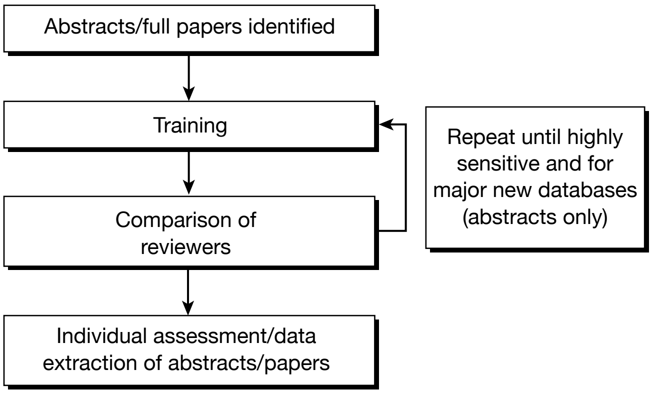

Abstracts/full papers identified
Training
Comparison of reviewers
Individual assessment/data extraction of abstracts/papers
Repeat until highly sensitive and for major new databases (abstracts only)

FIGURE 4 Process of training.

Rigorous training ensured high reliability of the screening process. To quantify this, the efficacy of training was assessed by calculating the area under the curve (AUC) devised from the receiver operating characteristic curve (ROC). The AUC was calculated against a 'gold standard' of exact matches arrived at by consensus. Having a sensitive process was considered more appropriate than overall agreement, with a focus on over-including (where in doubt in the early stages) being important to avoid missing key studies.

Three reviewers undertook abstract screening (AS, GG and KH). After the triplicate screening of 750 abstracts (three sets of 250) from MEDLINE, the ROCs were calculated, generating AUC scores: AS = 0.865, GG = 0.954, KH = 0.970. Training was repeated for PsycINFO, and 750 triplicate-screened abstracts generated AUC scores: AS = 0.88, GG = 0.92, KH = 0.90. Five reviewers undertook the initial screening of the full papers (AS, GG, KH, MR and CS). Training was carried out with 20 articles and reviewed independently. Consensus was achieved through discussion of included and excluded studies. Then a subsequent set of 20 studies were reviewed independently and the sensitivity of all reviewers was 100%. Data extraction and quality assessment were undertaken by three reviewers (GG, NC and RI). Training was carried out on two sets of 20 papers, giving AUC scores of GG = 0.823, NC = 0.802 and RI = 0.790.

Data extraction

The final extraction stage was carried out using a series of forms (see *Appendix 2*). These forms were circulated to all members of the study management team for comment and changes were implemented accordingly. As with each stage of the reviewing process, a training phase was completed. The data extraction was comprehensive because of the wide-ranging and diverse nature of the articles selected. Items for data extraction were selected to be as inclusive as possible; the details of each included study were captured under the following headings:

1. population and design (data forms 2 and 3)
2. mode description (data form 4)
3. measure description (data form 5)
4. comparison (data form 6).

Every paper reviewed had one form describing the setting and design of the study and its overall quality. For the other data forms, variable numbers were completed depending on the number of modes and measures compared. These were then linked using the unique study ID number.

Modes were put into a general categorisation, as well as classified by their mode features. The mode features were based on the theoretical framework developed in *Chapter 2* and additional features indicated as possibly related to response differences in the literature. The four main features from the theoretical framework were:

- administration (self or interviewer)
- telephone contact
- computerisation
- sensory stimuli (auditory, visual or both).

The first mode feature of administration is relatively self-explanatory. Modes in which an interviewer was recorded and then either played down the telephone, on video or on a computer are still classified as self-administered, as the control of the interview is with the respondents; for example, they can pause and play or stop at will.

The use of telephone could be by an interviewer or via an automated dial-up service for administration. The use of a computer can be in the form of a CATI, a CAPI or computer-based self-administration, such as a disk by post or a web survey. Sensory stimuli are coded on the basis of having purely auditory stimuli, such as simple telephone and face-to-face interviews; purely visual stimuli such as paper-based questionnaires or simple web surveys; or modes that combine both, such as face-to-face interviews with use of prompts such as flash cards or web-based surveys with a video/audio component.

Other features were coded to be tested for inclusion in the model. These related to the perceived legitimacy, such as how the measure was delivered to the respondent. This could be by telephone, in person or via the post/e-mail/web. Although the majority of telephone and face-to-face administered modes would have the same delivery as administration, for some studies these will be different, for example more laboratory-based studies in which all modes are introduced in person, but may still be completed as self-complete questionnaires or on a computer.

A number of other factors related to perceived anonymity, such as the mode of response provided, whether or not others were present during completion and whether or not anonymity was specifically protected. The ability to backtrack was also collected as a possible contributing factor to the level of cognitive burden.

For statistical data extraction, where standard deviation (SD) data were not presented, they were imputed from p-values, confidence intervals (CIs) or test statistics where available. Where information about scales, such as number of items, scoring, etc., was not provided in papers, the original source references for those studies were accessed for information.

Quality assessment

In order to assess the quality of the evidence contributing to this review, each paper was assessed for methodological quality. Assessing the quality of evidence becomes particularly challenging in the reviews of studies having diverse methodologies. In this particular review, RCTs were not necessarily expected, and so a more generic quality assessment tool was needed. Two tools were identified,[94,95] which provided quality checklists for studies other than RCTs.

Downs and Black[94] created a checklist for both randomised and non-randomised studies, focusing on health-care interventions. The checklist consisted of 27 items from five subscales:

1. *Reporting* Do the findings allow the reviewer to draw unbiased conclusions?
2. *External validity* Can the findings be generalised?
3. *Bias* Have potential biases been addressed and mentioned?
4. *Confounding* Have possible confounders been addressed and reported?
5. *Power* Could the findings be due to chance?

The tool, scored on a dichotomous scale, has good face validity, demonstrates inter-rater reliability and correlates well with an existing validated checklist, the Quality Index.[96] The checklist provides a detailed profile of both randomised and non-randomised studies.

Kmet *et al.*[95] took this process one step further by developing tools for both quantitative and qualitative research. The process, scored on a scale of zero to 2, evaluated the methodological choices and the clarity of reporting in relation to potential biases. However, the authors tested the checklist on only 10 articles, allowing a limited inter-rater reliability analysis.

The current tool was based upon the previous two checklists, with some modifications. The checklist of Downs and Black[94] is detailed containing 27 items, but is heavily weighted towards randomised designs. The Kmet *et al.*[95] checklist, although shorter at 14 items, focuses on intervention studies, which was not appropriate for this review. Therefore, it was necessary to create a checklist designed specifically for this present review that was more appropriate to both the methodological nature of the topic and the diverse literature base. The resulting checklist (see *Appendix 2*, datasheet 7) contained 18 items scored on three levels, yes (2), partial (1) and no (0), with three questions containing 'non-applicable' categories for specific study designs. Scores are summated across each item providing a percentage score, allowing consideration for the non-applicable items.

The piloting of quality assessment allowed testing of the inter-rater reliability. Both main reviewers (GG and RI) separately scored the quality of 20 papers included in the full data extraction phase. The scoring of each paper was carried out after the main descriptive and quantitative extraction of data from the papers. The detailed reading required for the data extraction process facilitated judgements of quality. As such, the checklist was quick and easy to complete, taking approximately 2 minutes per paper. Agreement between GG and RI was good, with κ-values on individual items ranging from 0.61 to 0.85. A paired-sample t-test on total scores demonstrated no significant differences between the reviewers (mean difference = 1.17, SD = 4.50, $p = 0.8$).

Publication bias

The conceptual framework for publication bias being based on journals and investigators not wanting to publish 'negative' studies is unlikely to apply in the case of mode comparison studies. The consideration that two modes are the same or different is equally likely to be newsworthy. Therefore, it is more likely that gaps in publications are likely to appear due to methodological reasons rather than outcome (poorly designed studies) or sample size (too small studies).

Evidence synthesis

An overview of the studies identified will be presented descriptively highlighting the different mode features identified in the theory review. Those with appropriate data will be subjected to quantitative methods of synthesis using exploratory metaregression[97] to identify the association between mode features and differences in response.

The primary analysis based on three key summary statistics is calculated for each comparison. These are:

- the absolute difference between the means (standardised) of the two modes
- the ratio of the largest to the smallest variance of the two modes
- the ES (absolute mean difference/SD) between two modes.

This allows for separate consideration of the accuracy and precision of the measures collected by the two modes as well as the more usual ES which combines both. For the first analysis, the mean differences need to be standardised to allow for measures on different scales to be combined. Using the highest and lowest possible scores on each scale, these were standardised to a 0–100 scale.

$$\text{Standardised score} = \frac{(\text{actual score} - \text{minimum value})}{(\text{maximum value} - \text{minimum value})} \qquad \text{[Equation 1]}$$

Where the average scores per item are used in summary statistics, the minimum and maximum values per item were used to standardise. The absolute value of the difference is used as, when combining many different outcomes, the direction of difference is meaningless.

For the second analysis the ratio of the two variances is already on a standardised scale as the largest variance is being presented as a proportion of the smallest. Similarly, the ES is a standardised statistic with the absolute mean difference expressed as a proportion of the SD. The pooled SD from the two modes was used in the calculation of the ES.

Between- and within-subject studies were analysed together, controlling for the study design. Analysis was conducted at two levels to account for clustering of comparisons within a study.[98] This allowed for study-level characteristics, measures characteristics and mode features to be considered together. The modelling approach assessed the four main mode features from the theoretical review, then tested the addition of other candidate features and then assessed model fit including other possible moderators of effect and identified interaction. Studies were categorised whether or not they were designed to show a difference on a mode feature. For example, this meant that a web versus a postal survey would have been coded as no difference on the feature of administration, whereas a web survey versus a telephone interview would have been coded as showing a difference. These differences were then used as explanatory variables in the models.

Sensitivity analysis explored the impact of weighting by quality scores (rather than using as an explanatory variable), as well as weighting by functions of the sample size and the pooled SD.

Statistical methods for individual within-group comparison studies for two methods of measuring the same entity have been debated extensively. This is particularly so in the field of clinical measurement where two clinical tools (e.g. thermometers) are compared on the same patients.[99–101] These techniques have varied from relatively simple methods for assessing accuracy and precision of instruments (e.g. limits of agreement and Bland–Altman plots[100]) to more complex modelling (e.g. structural equation modelling). Williamson *et al.*[102] developed two approaches to estimating combined limits of agreement[102] and the Mantel–Haenszel approach is presented for the two most frequently occurring scales, the Short Form questionnaire-36 items (SF-36) and MMPI. Studies that are between-group comparisons of these two sets of measures are also subjected to a standard random-effects meta-analysis. The original proposal was to undertake a review of the differences between studies of a single mode using SF-36; however, this was replaced with the meta-analysis above as being more appropriate given the number of studies identified which directly compared two modes using the SF-36. The MMPI was added owing to the number of studies reporting this outcome.

Analysis was undertaken using SPSS 14.0 (SPSS Inc., Chicago, IL, USA), MLwiN 1.1 (Centre for Multilevel Modelling, University of Bristol, Bristol, UK) and RevMan 5 (The Cochrane Collaboration, The Nordic Cochrane Centre, Copenhagen, Denmark).

Changes from original proposal

A number of minor changes were made to the original proposal, the search strategy was developed and refined from that in the original proposal when the theoretical review suggested that it was simplistic to simple categorise by crude mode and the training plan to ensure that individual reviewers was developed to incorporate all stages of review instead of simply the data extraction phase as stated. This was undertaken on a slightly smaller number of papers (20 rather than 25) than originally stated as agreement was good and individuals had already received considerable training in earlier phases. The major change was that the review of single-mode studies for SF-36 was replaced by a more detailed analysis of the mode comparison studies identified for that measure and also the MMPI. This decision was based on the numbers of studies identified.

This study is reported in accordance with reporting standards for systemic reviews and the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist[103] is included in *Appendix 4*.

Chapter 4

Results

The search strategy produced a total of 63,305 citations downloaded from the various databases, of which 39,253 were unique (*Figure 5*). These articles had their titles and abstracts reviewed, with 2156 articles being selected for retrieval in full. The full articles were then screened prior to detailed data extraction. The process excluded 1559 papers (see *Table 4* for details).

Studies excluded from the review

Table 4 shows the number of papers excluded from initial screening of the full 2156 papers and the reasons for their exclusion.

The most common reason for exclusion (44%) was that the paper did not contain a mode comparison. A number of studies (12%) described use of multiple modes of data collection; however, these were for different outcomes often measured at different time points. The next

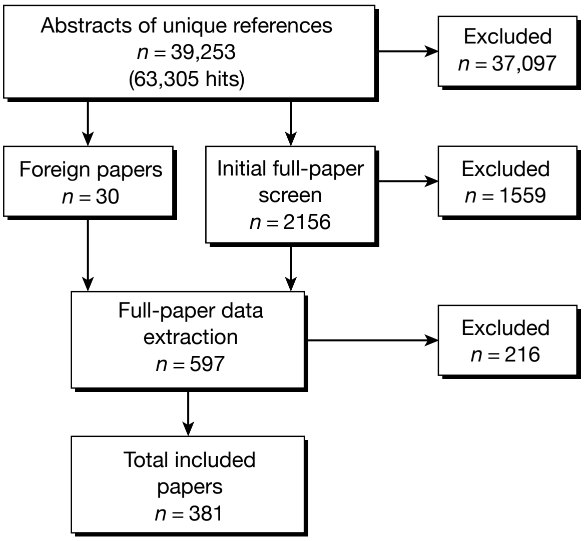


FIGURE 5 Flow diagram of study identification.

TABLE 4 Reasons for exclusion from the initial screen of the full paper

Reason	No.
No mode comparison	691
Mode comparison, but not comparing the same construct	91
Comparison of different judges	458
Measuring or comparing a behavioural construct only	230
Review (not primary study)	89
Total number of papers excluded at first stage	*1559*

most common reason (29%) was that the article referred to a comparison of two different judges, the most common of these being clinical diagnostic interviews for psychiatric disorders. As this incorporation of a second individual's judgement into one mode could invalidate the comparison, all structured clinical interviews have been excluded. The next largest group (15%) was that of papers that compared a behavioural construct only. These papers focused mainly on sensitive behaviours, such as smoking, sexual behaviour and drug taking. All of these papers were retrieved at abstract stage to be checked for any subjective component being reported, even when the main focus of the study was on measuring behaviour. Papers which solely focused on behaviour were excluded at this stage, whereas those that included some subjective elements were retained (e.g. being scared by your level of drinking would be included but the amount of alcohol drunk would not).

Of the 597 articles for which data extraction was undertaken, a total of 216 were also excluded (*Table 5*).

The most common reasons were that the construct being compared was not subjective (36%) or that it was judged by two different individuals (36%) (e.g. patient and clinician or parent and child). The next most common was if the paper contained no mode comparison (18%). This commonly occurred in studies in which there were two modes of data collection but no common data collected through multiple modes and therefore no mode comparison. An additional 13 papers (6%) were excluded as they only reported response rates and had no information on the actual responses given**.**

Thirty foreign-language articles were retrieved in full on the basis of their English title and abstract (where available). These were then screened for inclusion and data extracted where appropriate by one of the main reviewers (GG or RI) and a translator. The languages included Chinese, Danish, Dutch, French, German, Japanese and Spanish. During this process, it was found that 10 papers were to be excluded. Five further papers (two in Slovenian, two Russian and one Czech) were unable to be translated owing to the unavailability of a translator.

Description of included studies

Studies from 381 articles met the inclusion criteria for the review. There has been an increase in the number of published mode comparison studies over recent years (*Figure 6*). This increase in studies may represent many factors directly or indirectly linked to the methodology of mode comparison experiments. The first influence relates directly to the increase in technological options available to the survey researcher. However, direct factors such as the increase in the number of journals, particularly those that are electronic only, have led to a general increase in publications levels.

TABLE 5 Reasons for exclusion at the data extraction stage

Reason	No.
No mode comparison	39
Comparison of different judges	79
Measuring or comparing a behavioural construct only	79
Focuses on response rates only	13
Review	1
Unable to translate	5
Total number of papers excluded	*216*

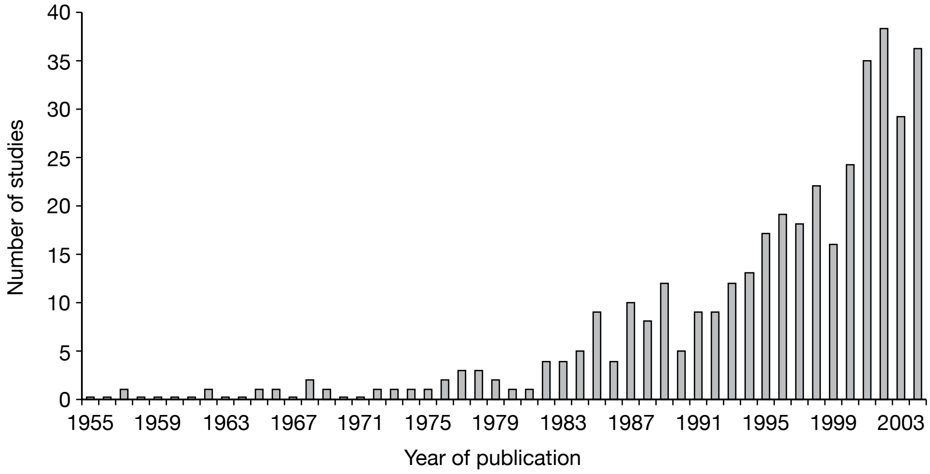


FIGURE 6 Number of mode comparison studies ($n = 381$) included by publication date.

Source of publication

Data were collected on the subject area in which the mode comparison was carried out. Most mode comparison studies were published in the area of health ($n = 201$, 53%). The next largest area of study for mode comparisons was psychology ($n = 86$, 23%) and social sciences ($n = 55$, 14%). The rest of the studies were focused on business ($n = 16$, 4%), statistics ($n = 14$, 4%) and education ($n = 9$, 2%).

Country and language of data collection

The review was not restricted by location of study or language, *Table 6* shows the distribution of the study locations. A large proportion of the studies were carried out in North America ($n = 236$, 62%), with 112 (29%) studies being carried out in Europe and 38 (10%) of those were from the UK.

The language of data collection was predominantly English ($n = 274$, 72%), although this was mostly inferred as it was clearly stated in only 30 (8%) these papers. The other languages used were predominantly European in origin, with French, German, Dutch and Spanish being the most frequent.

Study design

Studies were categorised based on the incorporated study design, either within subjects or between subjects. In total, 52% of studies ($n = 200$) were designed to provide a between-group comparison of modes, whereas 47% ($n = 180$) were within-group comparisons. Studies that were crossover by design have been included in the grouping in which they provided data for comparison (predominantly within groups).

Data were collected on whether the studies randomised either the mode an individual received (between-group studies) or the order in which modes were received (within-group studies). In total, 147 studies (39%) had used randomisation, with a higher proportion of between-group studies ($n = 83$, 42%) than within-group studies ($n = 64$, 36%) using this form of allocation. Studies which did not use randomisation used other forms of allocation such as drawing samples

TABLE 6 Geographical distribution of studies

Country	No. of studies
USA	201
Canada	36
UK	38
Germany	19
Australia	13
Netherlands	11
France	9
Sweden	7
Denmark	6
Spain	6
Norway	5
Switzerland	5
Belgium	2
Israel	2
Turkey	2
Austria	1
Brazil	1
China	1
Croatia	1
Finland	1
Hong Kong	1
Ireland	1
Italy	1
Japan	1
Mexico	1
Unknown	14
Total	*386*[a]

[a] Three studies were carried out in two countries and one study in three. Total number of studies = 381.

from separate sampling frameworks (e.g. separate population surveys[104] in between-group studies and systematic allocation (e.g. alternating[105,106]) for within-group studies. A relatively large number of within-group studies presented the modes under evaluation in exactly the same order to all participants ($n=95$, 53%).[107,108]

The 381 papers included in the review described 489 different samples. Some studies compared response on samples derived from two different sources (e.g. online survey panel compared with random-digit dialling). The methods for sampling demonstrated a dominance of two distinctly different approaches either by convenience ($n=155$, 32%) or targeting a specific group of participants, for example on a clinic list ($n=257$, 53%).

The measurement of study quality

The quality of every study included in the review was assessed utilising an 18-item tool specifically designed for the present review. The tool measures quality of quantitative studies irrespective of study design. Overall scores were generally high (*Figure 7*). However, certain items showed a higher percentage of poor ratings than the others. These were the items relating to clear descriptions of participants (22% poor), group allocation (50% poor), appropriate consideration given to the impact of timing of data collection (27% poor) and reporting of variances for results (35% poor). However other items had extremely high scores such as having a clearly stated hypothesis (89% good), the study design described and appropriate (83% good) and the conclusions supported by the results (81% good).

Measures used

In total, the 381 papers provided 1282 measure descriptions. Thirty per cent of studies considered only a single measure, with one study comparing 21 different measures (*Figure 8*). The term measure did not relate solely to one tool, but to the subscales within the measure, for example a study that reported using all subscales of the SF-36 would represent eight measures.

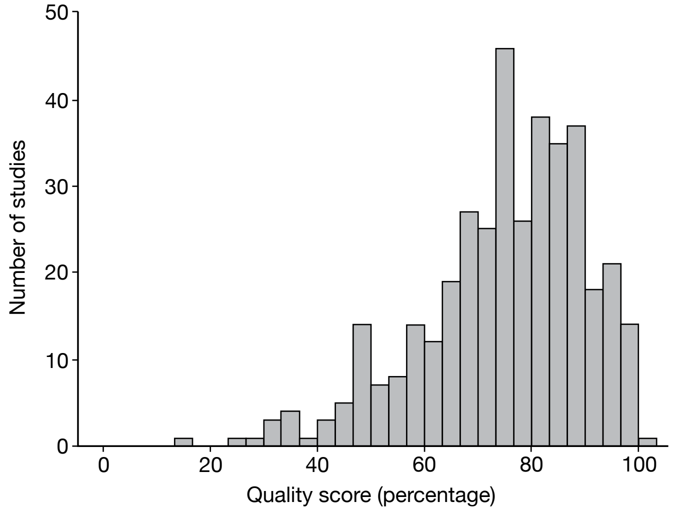


FIGURE 7 Frequency distribution of percentage quality scores for included studies.

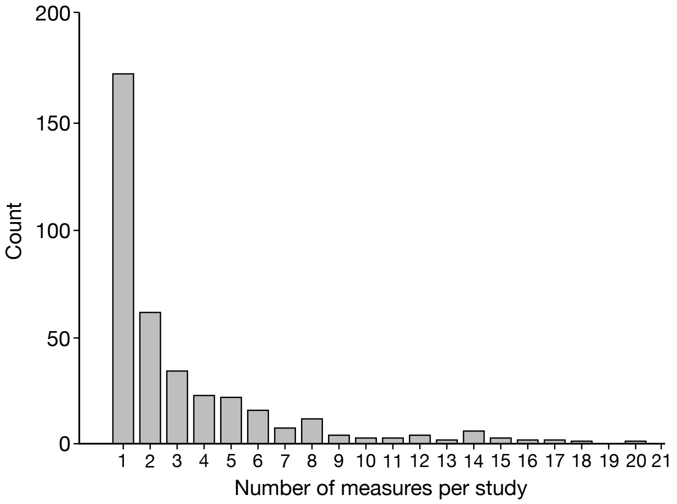


FIGURE 8 The number of studies by number of measures reported.

Each measure described was categorised as whether it concerned a health-related area or not. Measures such as QoL symptoms, as well as those relating to general mental well-being (anxiety, etc.) were classified as health and those measuring societal attitudes, personality and willingness to pay were classified as non-health. Of the 1282 measures described, 733 (57%) were classified as being health related.

To examine further the type of constructs measured, the measures were categorised based upon the psychological construct being measured. Studies measuring personality (*n* = 257, 20%) and specific aspects and dimensions of QoL (*n* = 215, 17%) were the most common. The most frequently occurring scales were the SF-36 (17 studies) and the MMPI (nine studies), which have 8 and 14 subscales, respectively, and therefore dominate the QoL and personality assessment categories. It should also be acknowledged that the categorisation is as driven by the description from the scale developers, and for some scales there may be little difference, for example, between the types of measures which have been classified as QoL and those classified as functional health status.

Modes evaluated

In total, the 381 papers described 801 modes. All studies provided a comparison between at least two modes (because of the inclusion criteria); however, some studies compared more, with 35 (9%) comparing three modes and two studies (1%) comparing four modes.

Each mode can be roughly categorised into one of four groups by main delivery method. These can be considered to be:

- computer (including web)
- paper
- telephone
- in person (face to face).

Although the features identified in the theoretical review cut across these categories, all the comparisons identified are between rather than within these categories. The total numbers of papers (and comparisons) by comparison group are given in *Table 7*.

As well as the relatively simplistic categorisation above, a more detailed level of information was obtained relating to specific features of the survey mode. This stratification was defined by the work of Tourangeau *et al.*[8] and discussed in *Chapter 2*. This theoretical framework defines four

TABLE 7 Number of comparisons and studies by comparison group

Comparison	No. of studies	No. of comparisons	Comparisons per study: mean (median)	Range
Computer vs paper	161	665	4.1 (2)	1 to 23
Computer vs telephone	12	17	1.4 (1)	1 to 3
Computer vs person	22	50	2.3 (2)	1 to 11
Paper vs telephone	74	280	3.8 (2)	1 to 36
Paper vs person	106	367	3.5 (2)	1 to 24
Telephone vs person	52	143	2.8 (1.5)	1 to 11
Overall	383[a]	1522	4.0 (2)	1 to 36

[a] Some papers appear in more than one category.

main mode features (administration, telephone contact, computerisation and sensory stimuli). Additional mode features related to other potential mediating factors are also explored whether or not they explain variation over and above that explained by the four main features.

Four main mode features

Of the total number of comparisons made, 667 (44%) involved a comparison between administration by an interviewer and self-completion. Telephone contact was one of the differences between modes for 440 (29%) of the comparisons (*Table 8*). Computers were incorporated in data collection of one mode in 803 (53%) comparisons. There was a difference in the main sensory stimuli in 714 (47%) comparisons.

Other possible mode features

Other features that were considered were the methods of delivery and response for the measure, whether or not the measure was completed 'online' (i.e. inputted through a technological device, which is connected to another technological device in 'real time', such as a telephone connected to another telephone or computer), who was physically present during completion (interviewer/other), the degree of anonymity of the process and the ability to backtrack through questions, and whether the response was oral, written or by means of electronics (e.g. pushing buttons).

The presence of others (not including the interviewer/researcher) and the ability to backtrack through a questionnaire were only explicitly mentioned in 6% of comparisons. Although reported in more studies, the degree of anonymity was different in only 13 comparisons (1%). None of these three features is therefore included in further modelling.

Possible mediators

The key theoretical mediating factors within the model presented in *Chapter 2* are impersonality, legitimacy and cognitive burden. There were no direct measures of impersonality that are reported in the studies, and any indirect assessment is instead inferred from the description of the mode features above. The issue is similar for legitimacy, although information on the source of approach for a study was recorded in 350 papers (92%). However, the source was a public body

TABLE 8 Numbers of comparisons reporting a difference in mode features

Mode feature	Difference	No difference	Missing
Administration	667	855	0
Telephone	440	1082	0
Computer	803	719	0
Sensory stimuli	714	808	0
Delivery method	686	836	0
Presence of interviewer/researcher	672	850	0
Online/offline	523	999	0
Response method	1386	136	0
Presence of others	13	82	1427
Anonymity	13	1493	16
Ability to backtrack	11	83	1428

(university, hospital or other) in 331 cases (87%) and a private company in only 4%; therefore, this is not included in further analysis. The only additional consistently available information relating to cognitive burden was the number of items in a scale. Where this was not available from a paper it was gathered from elsewhere, giving information for 1456 (96%) comparisons. This is therefore included as a mediating factor in the meta-analysis. As the number of items per measure is highly skewed with a small number of outcomes having very large numbers of items, it was categorised by four percentile groups (*Table 9*).

An additional factor suggested in some reviews for technology-assisted data collection is timing of the study. When a technology is first introduced and is novel to the individuals within the study, greater differences may occur than once familiarisation has taken place. Date of data collection was poorly reported in studies, with 295 (77%) studies giving no indication of when their sample was recruited or data collected. Therefore, date of publication of the paper is used as an approximation to this. This distribution was highly skewed and, therefore, the data have been transformed.

Assessment of mode effects on systematic bias

Of the 1522 comparisons, 977 gave information to enable the calculation of a standardised mean difference. The mean within each mode was standardised and then the absolute mean difference between the two means taken as the summary statistic for this analysis. As this gives rise to an exponential distribution, the log of the absolute difference (plus 1) was taken for further analysis (*Figure 9*). This gives rise to a distribution that is left truncated at zero, but which, given the sample size, can be taken as normal for further analysis. This summary statistic captures the magnitude of differences between two modes on a standardised scale, so values can be interpreted as percentage differences.

Only 53% of studies contribute to this analysis; however, these represent 64% of the comparisons as those studies that report more comparisons are also reporting the data needed to calculate this summary statistic. As might be expected for this type of review, the level of clustering of outcome within studies overall is high [intracluster correlation (ICC) = 0.37], with studies considering within-person comparison of modes having a higher ICC (0.62) than between-group comparison studies (0.15). The ICC gives an indication of how similar the results are across the different outcomes measured within the same study.

A two-level linear regression model was fitted to the log of the absolute mean difference. The first model (*Box 2*) was fitted with the four main mode features representing the theoretical framework. Then the addition of other possible features was tested in model 2. The addition of date of publication as a mediating factor and interactions with the main mode features is included in model 3, as well as testing for the effect of study design. Model 4 is based on the

TABLE 9 Percentile groups for number of items within each measure

No. of items	*n*	%
1	369	24.2
2–5	377	24.8
6–18	335	22.0
19+	375	24.6
Missing	66	4.3
Total	*1522*	*100.0*

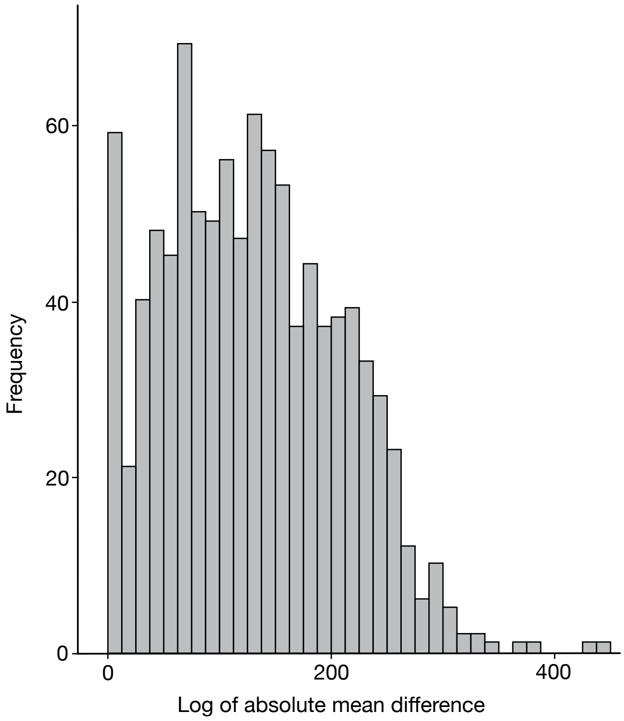


FIGURE 9 Histogram of logarithm of the absolute mean difference.

BOX 2 Summary of models fitted

Model 1: features from theoretical framework

Model 2: model 1 + suggested other features

Model 3: model 2 + date of publication and specified interactions

Model 4: model 1 + anything significant from models 2 and 3 + cognitive burden (no. of items)

subset of comparisons with data on the number of items per measure. Each mode feature was coded to represent whether the two modes compared showed a difference on that feature or not, therefore a comparison of a face-to face interview where the questions were read out loud and a telephone interview would have no difference in terms of sensory stimuli (both auditory), method of administration (both interviewer) or response (both verbal), but would show a difference in terms of use of a telephone and being online.

Fitting the model with absolute mean difference between the two mode features, we observed that, of the four main mode features, differences in administration (interviewer vs self) are highly significantly associated with larger differences between modes (*Table 10*). Differences in sensory stimuli are also significant, whereas the use of a computer or telephone has no impact on the magnitude of the difference between modes. On testing the additional possible features of mode (model 2), only the method of delivery approached significance and was, therefore, retained for further models. Model 3 shows that the date of publication is not associated with the magnitude of the difference and there are no significant interactions with the features associated with emerging technology (computer, telephone, sensory stimuli and delivery). The design of the study also had no impact on the model. Model 4 is fitted to the 941 comparisons in which data on the number of items within the measure are available. This shows a significant main effect with

TABLE 10 Two-level regression models for absolute mean difference between two modes

	Model 1: $n = 977$		Model 2: $n = 977$		Model 3: $n = 977$		Model 4: $n = 941$	
Variable	**B (SE)**	***p*-value**	**B (SE)**	***p*-value**	**B (SE)**	***p*-value**	**B (SE)**	***p*-value**
Administration	**0.69 (0.19)**	**< 0.001**	**0.86 (0.28)**	**< 0.001**	**0.69 (0.19)**	**< 0.001**	**0.67 (0.19)**	**< 0.001**
Sensory stimuli	**–0.44 (0.18)**	**0.01**	**–0.37 (0.19)**	**0.05**	–0.30 (0.26)	0.29	**–0.43 (0.18)**	**0.02**
Computer	–0.10 (0.11)	0.91	0.10 (0.14)	0.49	0.35 (0.27)	0.18	0.04 (0.11)	0.70
Telephone	0.09 (0.08)	0.29	–0.17 (0.17)	0.30	0.02 (0.28)	0.94	–0.09 (0.10)	0.39
Delivery			**0.24 (0.12)**	**0.05**	**0.54 (0.22)**	**0.01**	**0.26 (0.10)**	**0.01**
Response			–0.19 (0.18)	0.29				
Online			0.07 (0.18)	0.75				
Presence of interviewer			–0.12 (0.24)	0.61				
Design					0.11 (0.08)	0.21		
Date of publication					0.21 (0.13)	0.11		
Date *by* sensory stimuli					–0.08 (0.11)	0.13		
Date *by* computer					–0.19 (0.12)	0.82		
Date *by* telephone					–0.03 (0.13)	0.44		
Date *by* delivery					–0.18 (0.11)	0.09		
No. of items								
1							**Ref.**	**0.01**
2–5							**–0.21 (0.10)**	
6–18							**–0.31 (0.10)**	
19+							**–0.28 (0.11)**	
	Variance		**Variance**		**Variance**		**Variance**	
Level 2	0.20 (0.03)	< 0.001	0.20 (0.03)	< 0.001	0.19 (0.03)	< 0.001	0.21 (0.03)	< 0.001
Level 1	0.38 (0.02)	< 0.001	0.37 (0.02)	< 0.001	0.37 (0.02)	< 0.001	0.37 (0.02)	< 0.001
–2LLH	2030.25	Ref.	2021.52	0.07	2016.48	0.03	n/a[a]	

B, regression coefficient; LLH, log-likelihood; n/a, not applicable; Ref., reference.

[a] Not comparable to the other –2LLHs.

Bold text indicates $p < 0.05$.

scales with more than one item associated with smaller differences between modes; however, there were no significant interactions with the mode features. This suggests that differences between modes reduce with increasing number of items and therefore cognitive burden.

Assessment of mode effects on precision (variability)

Of the 1522 comparisons, 910 (60%) gave information on the SD or variance for each mode. One paper was excluded from this analysis because of the exceptionally large differences between variances (in excess of 100) suggestive of typographical errors. A two-level linear regression model was fitted as for the standardised mean difference (*Table 11*).

None of the mode features was associated with the size of the ratio of variances. The only variable that was significant was the design of the study, with between-group studies having greater differences between variances than within-group designs. This is as would be expected. No interactions were tested, as none of the main effects was significant.

TABLE 11 Two-level regression models for ratio of the variances between two modes

	Model 1: $n = 910$		Model 2: $n = 910$		Model 3: $n = 910$		Model 4: $n = 888$	
Variable	B (SE)	*p*-value	B (SE)	*p*-value	B (SE)	*p*-value	B (SE)	*p*-value
Administration	0.18 (0.23)	0.44	0.28 (0.34)	0.40	0.07 (0.23)	0.75	0.08 (0.22)	0.74
Sensory stimuli	–0.15 (0.22)	0.48	–0.03 (0.24)	0.91	–0.07 (0.21)	0.73	–0.07 (0.21)	0.72
Computer	0.02 (0.13)	0.88	0.17 (0.16)	0.30	0.02 (0.13)	0.90	0.03 (0.13)	0.84
Telephone	–0.14 (0.11)	0.20	–0.30 (0.20)	0.13	–0.13 (0.11)	0.21	–0.11 (0.11)	0.29
Delivery			0.15 (0.16)	0.36				
Response			–0.33 (0.22)	0.14				
Online			–0.04 (0.22)	0.86				
Presence of interviewer			–0.05 (0.28)	0.85				
Design					**0.24 (0.10)**	**0.01**	**0.25 (0.10)**	**0.01**
Date of publication					0.04 (0.05)	0.42	0.04 (0.05)	0.43
No. of items								
1							Ref.	0.10
2–5							–0.28 (0.14)	
6–18							–0.32 (0.14)	
19+							–0.19 (0.15)	
	Variance		Variance		Variance		Variance	
Level 2	0.26 (0.05)	< 0.001	0.25 (0.05)	< 0.001	0.23 (0.04)	< 0.001	0.21 (0.04)	< 0.001
Level 1	0.57 (0.03)	< 0.001	0.56 (0.03)	< 0.001	0.57 (0.03)	< 0.001	0.56 (0.03)	< 0.001
–2LLH	2242.70	Ref.	2238.93	0.44[a]	2235.98	0.03[a]	n/a	

B, regression coefficient; LLH, log-likelihood; n/a, not applicable; Ref., reference.

[a] Compared with model 1.

Bold text indicates $p < 0.05$.

Assessment of mode effects on overall effect size

Data were available to calculate the ES for 912 comparisons (60%) (*Table 12*). The ES was calculated as the absolute difference between the means (raw) divided by the pooled SD.

Two-thirds of the ESs would be considered negligible (< 0.2). This was highly skewed and, therefore, this was transformed prior to analysis (*Figure 10*).

A series of two-level linear regression models were then fitted as for the absolute mean difference (*Table 13*). The feature of administration is highly significant across all models, indicating a greater effect of this on the magnitude of differences between modes. Differences in sensory stimuli are of borderline significance in most models. Both the design of the study and the date of publication were significantly associated with ES. There were significant interactions between date of publication and computer and telephone usage. The numbers of items was significantly associated with ES, with smaller ESs for scales longer than one item. There was a significant interaction between this and the use of a computer.

TABLE 12 Effect sizes in categories[109]

ES	No. (%)
0.0–0.1999	604 (66.2)
0.2–0.3999	176 (11.3)
0.4–0.5999	67 (7.3)
0.6–0.9999	50 (5.5)
1.0–1.9999	12 (1.3)
≥ 2.0 and greater	3 (0.3)

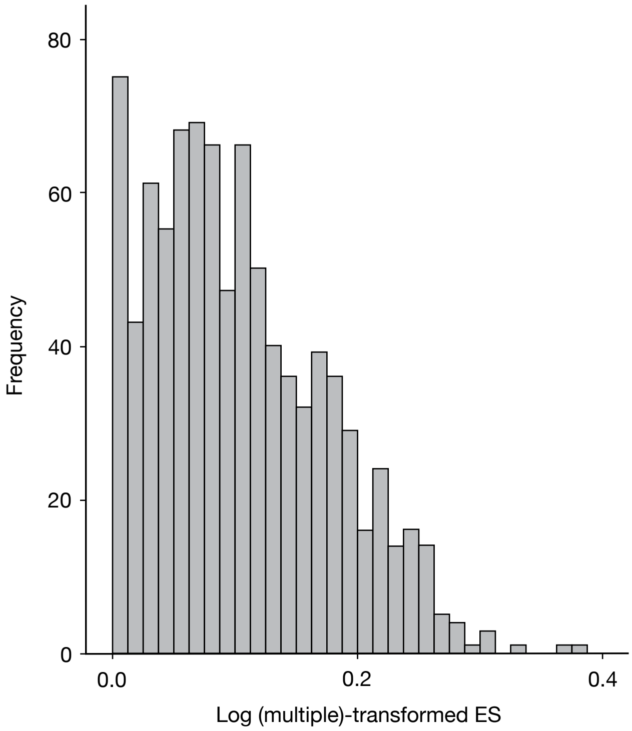


FIGURE 10 Distribution of transformed ES.

Interpretability of results

The greatest impact of mode features is on the systematic bias in responses rather than the variability of responses. If we were to take a hypothetical example for a measure, such as a subscale with two to five items from the SF-36 scored from 0 to 100, then the impact of the two significant variables 'administration' and 'sensory stimuli' on the absolute mean difference (systematic bias) is shown in *Table 14*, in terms of the predicted absolute mean differences.

This is what we would predict in terms of absolute mean difference if we were to design a factorial trial with two measurements carried out on each participant. However, if we want to relate this to mean difference (instead of absolute mean difference), we need to make some assumptions. It is reasonable to assume that, in the absence of any differences in mode or features causing biased responding, that the upper right-hand cell represents a half-normal distribution centred on zero. This relates to a normal distribution for differences with a mean of zero and an estimated SD of approximately '5'. The most commonly occurring combination of these two mode features is to have both a difference in administration and a difference in sensory stimuli, which, for a measure such as the SF-36, would result in an expected bias of 0.85 units, assuming no impact on the SD.

TABLE 13 Two-level regression models for ES between two modes

	Model 1: *n* = 912		Model 2: *n* = 912		Model 3: *n* = 912		Model 4: *n* = 888		Model 5: *n* = 888	
Variable	**B (SE)**	***p*-value**	**B (SE)**	***p*-value**	**B (SE)**	***p*-value**	**B (SE)**	***p*-value**	**B (SE)**	***p*-value**
Administration	**0.57 (0.20)**	**0.003**	**0.71 (0.30)**	**0.02**	**0.57 (0.19)**	**0.003**	**0.56 (0.19)**	**0.003**	**0.57 (0.19)**	**0.003**
Sensory stimuli	**–0.38 (0.19)**	**0.05**	–0.35 (0.20)	0.08	–0.27 (0.23)	0.23	**–0.39 (0.18)**	**0.03**	**–0.39 (0.18)**	**0.03**
Computer	0.14 (0.10)	0.16	0.24 (0.13)	0.06	**0.57 (0.18)**	**0.001**	**0.50 (0.15)**	**0.001**	–0.03 (0.25)	0.89
Telephone	0.12 (0.08)	0.14	0.13 (0.16)	0.40	**0.83 (0.27)**	**0.002**	**0.74 (0.25)**	**0.003**	**0.67 (0.25)**	**0.008**
Delivery			**0.36 (0.13)**	**0.007**	0.10 (0.19)	0.58	0.16 (0.10)	0.11	0.15 (0.10)	0.13
Response			–0.13 (0.17)	0.44						
Online			–0.29 (0.18)	0.10						
Presence of interviewer			–0.08 (0.25)	0.74						
Design					**0.19 (0.09)**	**0.04**	**0.20 (0.09)**	**0.03**	**0.22 (0.09)**	**0.02**
Date of publication					**0.26 (0.11)**	**0.01**	**0.20 (0.08)**	**0.009**	**0.18 (0.08)**	**0.02**
Date *by* sensory stimuli					–0.07 (0.08)	0.37				
Date *by* computer					**–0.28 (0.09)**	**0.003**	**–0.23 (0.08)**	**0.003**	**0.21 (0.08)**	**0.009**
Date *by* telephone					**–0.43 (0.13)**	**0.001**	**–0.37 (0.11)**	**0.001**	**–0.33 (0.11)**	**0.004**
Date *by* delivery					0.03 (0.10)	0.93				
No. of items										
1							**Ref.**	**0.002**	**Ref.**	**0.001**
2–5							**–0.26 (0.11)**		**–0.46 (0.15)**	
6–18							**–0.32 (0.11)**		**–0.62 (0.15)**	
19+							**–0.13 (0.12)**		**–0.44 (0.17)**	
Computer by no. of items										
1									**Ref.**	**0.02**
2–5									**0.37 (0.21)**	
6–18									**0.58 (0.21)**	
19+									**0.59 (0.22)**	
	Variance		**Variance**		**Variance**		**Variance**			
Level 2	0.28 (0.04)	< 0.001	0.29 (0.04)	< 0.001	0.26 (0.04)	< 0.001	0.25 (0.04)	< 0.001	0.25 (0.04)	< 0.001
Level 1	0.26 (0.01)	< 0.001	0.26 (0.01)	< 0.001	0.26 (0.01)	< 0.001	0.25 (0.01)	< 0.001	0.25 (0.01)	< 0.001
–2LLH	1656.52	Ref.	1648.25	0.08	1632.93	0.02	1565.82	Ref.	1556.36	0.02

B, regression coefficient; LLH, log-likelihood; n/a, not applicable; Ref., reference.

a Not comparable to the other –2LLHs.

TABLE 14 Predicted absolute mean differences from Model 4

	Difference in administration	
	No	Yes
Difference in sensory stimuli		
No	2.07	5.01
Yes	1.01	2.92

Meta-analysis of Short Form questionnaire-36 items

The most frequently occurring individual outcome measure within the studies included was the SF-36 health survey.[110] The SF-36 consists of eight aggregate scale scores. Each scale is directly transformed into a 0–100 scale on the assumption that each contributing item carries equal weight. The eight scales are vitality, physical functioning, bodily pain, general health, role physical, role emotional, role mental and mental health.

Seventeen studies[104,111–126] published between 1994 and 2003 used SF-36. Not all studies reported all subscales. The impact of the different modes of using SF-36 was assessed using weighted pooled measures of agreement for within-subject comparisons and random-effects meta-analysis for between-subject comparison. There were seven studies[104,111,112,114–117] that provided between-subject comparisons only, eight studies[119–126] that provided within-subject comparisons only and two studies[113,118] that contributed data to both analyses. *Table 15* summarises the information available from each study.

Between-subject comparisons

Eight studies[104,112,114–118] had some data available that could contribute to the meta-analysis. One of these (Amodei *et al.*[111]) was a comparison of an interview in which the interviewer asked the questions and recorded the response to one in which the interviewer asked the questions and the responder confidentially recorded their own response.[111] This mode comparison does not reflect a difference on one of the four mode features and, therefore, has not been included in the subsequent analysis. One of the crossover studies (Lyons *et al.*[113]) provided only mean scores at the first time point and, therefore, could not be included in this analysis.[113]

Within-subject comparisons

There was a greater variety in the statistical approaches taken to analysis in the within-subject studies, and the data presented that could contribute to the pooled analysis were limited. Studies that did not give information on mean differences and SDs tended to report correlations. The available studies have been combined to give pooled estimators of mean difference with 95% CIs and pooled limits of agreement.

Mode feature: computer

Only two of the between-subjects studies contributed to the analysis of the computerisation mode feature.[114,115] The results of the meta-analysis for each subscale of the SF-36 can be seen in *Figures 11–18* (forest plots in order of magnitude of pooled difference).

Role emotional, social functioning and mental health (see *Figures 11–13*) all suggest that significantly higher scores are achieved with computers than without, with mean differences of between four and eight points on the scale. It should be noted that as the Perkins and Sanson-Fischer study[114] is 10 times the size of the Saleh *et al.* study,[115] it dominates the pooled estimator.

TABLE 15 Included studies using SF-36

Paper	Country	Population	Design	Comp	Adm	Tel	Sens	*Data availability*
Between-subject comparisons								
Amodei (2003)[111]	USA	Primary care (non-psychiatric)	Randomised trial	n	n	n	n	Data available, but no differences on mode features
Bowling (1999)[104]	UK	General population	Two separate surveys	n	y	n	y	Yes
Jones (2001)[112]	USA	Outpatients	Randomised trial with crossover for non-responders	n	y	y	y	Data taken prior to crossover
Lyons (1999)[113]	UK	Outpatients	Randomised crossover	n	y	n	y	Data taken prior to crossover (no SDs)
Perkins (1998)[114]	Australia	General population	Randomised trial	y	y	y	y	Yes
Saleh (2002)[115]	USA	Outpatients (orthopaedics)	Non-randomised trial	y	n	n	n	Yes
Unruh (2003)[116]	USA	Haemodialysis patients	Randomised trial	n	y	n	y	Yes
van Campen (1998)[117]	Netherlands	Patients with chronic illnesses	Randomised trial	n	y	y	y	No data
Weinberger (1996)[118]	USA	Patients with chronic illnesses	Randomised crossover	n	y	y	y	Data taken prior to crossover
Within-subject comparisons								
Abdoh (2001)[119]	USA	Patients	Unclear	y	n	n	y	No data
Bliven (2001)[120]	USA	Outpatients (cardiology)	Randomised crossover	y	n	n	n	No data
Caro (2001)[121]	Canada	Outpatients (asthma)	Alternating crossover	y	n	n	n	No data
Lyons (1999)[113]	UK	Outpatients	Randomised crossover	n	y	n	y	Data taken combining order groups
Molitor (2001)[122]	USA	People living in transitional housing	Sequential crossover	n	n	n	y	No data
Revicki (1997)[123]	USA	Patients with bipolar disorder	Randomised crossover	n	n	y	n	Data taken combining order groups
Ryan (2002)[124]	Australia	Healthy adults	Randomised crossover	y	n	*n*	n	Data taken combining order groups
Weinberger (1994)[125]	USA	General medical care, aged 65+ years		n	n	y	n	No data
Weinberger (1996)[118]	USA	Patients with chronic illnesses	Randomised crossover	*n*	y	y	y	Data taken from comparison of first crossover
Wilson (2002)[126]	UK	Outpatients (rheumatology)	Crossover	y	n	n	n	No data

Adm, administration; Comp, computerisation; n, no; Sens, sensory stimuli; Tel, telephone; y, yes.
Shaded cells indicate that these studies contributed to the analysis of that mode feature.

Only one of the within-subjects studies[124] provided data on this mode feature. The results for the Ryan *et al.* study are given in *Table 16*.

The only outcome for which there was a significant difference was for 'social functioning', with higher scores for those not using a computer. This is contrary to the findings in the

	Computerised			Not computerised				
Study or subgroup	Mean	SD	Total	Mean	SD	Total	Weight (%)	Mean difference IV, Random, 95% CI
Perkins 1998[114]	88.76	27.99	421	80.36	33.47	418	95.3	8.40 (4.22 to 12.58)
Saleh 2002[115]	56.9	44.2	41	55.6	42.7	42	4.7	1.30 (–17.40 to 20.00)
Total (95% CI)			**462**			**460**	**100.0**	**8.06 (3.99 to 12.14)**

Heterogeneity: $\tau^2 = 0.00$; $\chi^2 = 0.53$, df = 1 ($p = 0.47$); $I^2 = 0\%$
Test for overall effect: $z = 3.88$ ($p = 0.0001$)

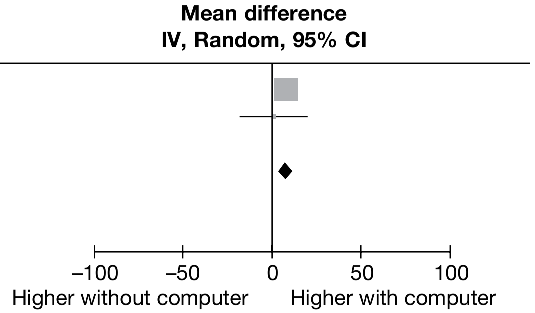


FIGURE 11 Meta-analysis SF-36 – role emotional*. *, $p < 0.05$.

	Computerised			Not computerised				
Study or subgroup	Mean	SD	Total	Mean	SD	Total	Weight (%)	Mean difference IV, Random, 95% CI
Perkins 1998[114]	89.25	22.6	421	84.59	21.2	418	94.3	4.66 (1.70 to 7.62)
Saleh 2002[115]	65.4	27.4	39	65.3	28.9	45	5.7	0.10 (–11.95 to 12.15)
Total (95% CI)			**460**			**463**	**100.0**	**4.40 (1.52 to 7.28)**

Heterogeneity: $\tau^2 = 0.00$; $\chi^2 = 0.52$, df = 1 ($p = 0.47$); $I^2 = 0\%$
Test for overall effect: $z = 3.00$ ($p = 0.003$)

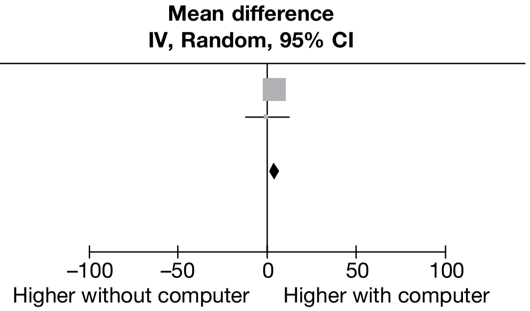


FIGURE 12 Meta-analysis SF-36 – social functioning*. *, $p < 0.05$.

	Computerised			Not computerised				
Study or subgroup	Mean	SD	Total	Mean	SD	Total	Weight (%)	Mean difference IV, Random, 95% CI
Perkins 1998[114]	80.38	17.42	421	76.33	16.28	418	93.4	4.05 (1.77 to 6.33)
Saleh 2002[115]	27.9	23	39	25.7	16.1	45	6.6	2.20 (–6.42 to 10.82)
Total (95% CI)			**460**			**463**	**100.0**	**3.93 (1.72 to 6.13)**

Heterogeneity: $\tau^2 = 0.00$; $\chi^2 = 0.17$, df = 1 ($p = 0.68$); $I^2 = 0\%$
Test for overall effect: $z = 3.49$ ($p = 0.0005$)

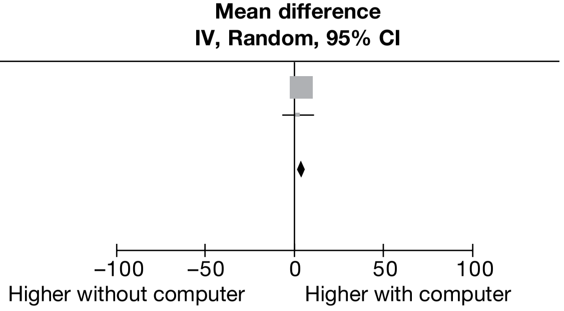


FIGURE 13 Meta-analysis SF-36 – mental health*. *, $p < 0.05$.

	Computerised			Not computerised				
Study or subgroup	Mean	SD	Total	Mean	SD	Total	Weight (%)	Mean difference IV, Random, 95% CI
Perkins 1998[114]	80.34	35.7	421	76.77	36.44	418	92.4	3.57 (–1.31 to 8.45)
Saleh 2002[115]	27.4	40.2	41	30.9	39.9	44	7.6	–3.50 (–20.54 to 13.54)
Total (95% CI)			**462**			**462**	**100.0**	**3.03 (–1.66 to 7.73)**

Heterogeneity: $\tau^2 = 0.00$; $\chi^2 = 0.61$, df = 1 ($p = 0.43$); $I^2 = 0\%$
Test for overall effect: $z = 1.27$ ($p = 0.21$)

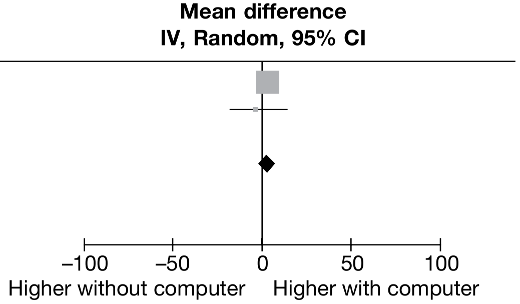


FIGURE 14 Meta-analysis SF-36 – role physical. *, $p < 0.05$.

	Computerised			Not computerised				
Study or subgroup	Mean	SD	Total	Mean	SD	Total	Weight (%)	Mean difference IV, Random, 95% CI
Perkins 1998[114]	63.22	22.78	421	60.57	21.48	418	88.7	2.65 (–0.35 to 5.65)
Saleh 2002[115]	48.7	22.4	39	49	15.8	45	11.3	–0.30 (–8.71 to 8.11)
Total (95% CI)			**460**			**463**	**100.0**	**2.32 (–0.50 to 5.14)**

Heterogeneity: $\tau^2 = 0.00$; $\chi^2 = 0.42$, df = 1 ($p = 0.52$); $I^2 = 0\%$
Test for overall effect: $z = 1.61$ ($p = 0.11$)

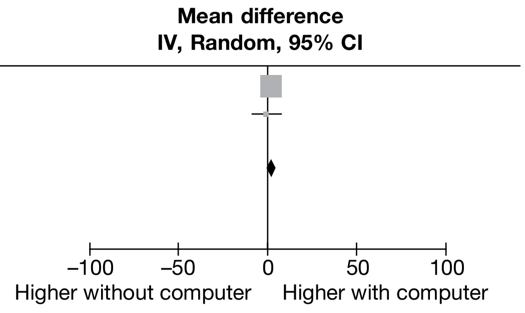


FIGURE 15 Meta-analysis SF-36 – vitality.

	Computerised			Not computerised				
Study or subgroup	Mean	SD	Total	Mean	SD	Total	Weight (%)	Mean difference IV, Random, 95% CI
Perkins 1998[114]	72.84	22.7	421	71.01	21.64	418	89.8	1.83 (–1.17 to 4.83)
Saleh 2002[115]	56	20.8	36	56.7	19.7	45	10.2	–0.70 (–9.60 to 8.20)
Total (95% CI)			**457**			**463**	**100.0**	**1.57 (–1.27 to 4.42)**

Heterogeneity: $\tau^2 = 0.00$; $\chi^2 = 0.28$, df = 1 ($p = 0.60$); $I^2 = 0\%$
Test for overall effect: $z = 1.08$ ($p = 0.28$)

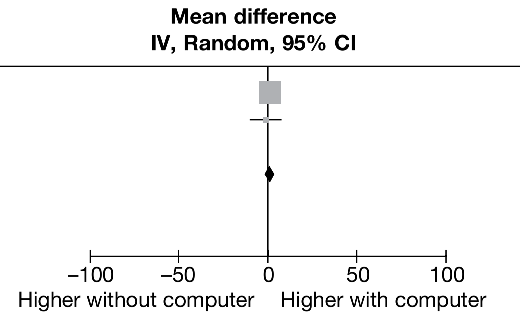


FIGURE 16 Meta-analysis SF-36 – general health perception.

	Computerised			Not computerised				
Study or subgroup	Mean	SD	Total	Mean	SD	Total	Weight (%)	Mean difference IV, Random, 95% CI
Perkins 1998[114]	82.64	24.16	421	81.35	22.8	418	90.5	1.29 (–1.89 to 4.47)
Saleh 2002[115]	40.5	19.5	41	38.9	26.3	44	9.5	1.60 (–8.20 to 11.40)
Total (95% CI)			**462**			**462**	**100.0**	**1.32 (–1.70 to 4.34)**

Heterogeneity: $\tau^2 = 0.00$; $\chi^2 = 0.00$, df = 1 ($p = 0.95$); $I^2 = 0\%$
Test for overall effect: $z = 0.86$ ($p = 0.39$)

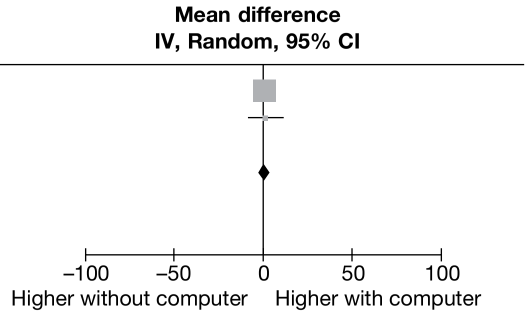


FIGURE 17 Meta-analysis SF-36 – physical functioning.

	Computerised			Not computerised				
Study or subgroup	Mean	SD	Total	Mean	SD	Total	Weight (%)	Mean difference IV, Random, 95% CI
Perkins 1998[114]	77.75	27.64	421	73.73	24.41	421	59.2	4.02 (0.50 to 7.54)
Saleh 2002[115]	36.2	17.6	39	40.1	17.4	44	40.8	–3.90 (–11.45 to 3.65)
Total (95% CI)			**460**			**465**	**100.0**	**0.79 (–6.84 to 8.42)**

Heterogeneity: $\tau^2 = 22.34$; $\chi^2 = 3.47$, df = 1 ($p = 0.06$); $I^2 = 71\%$
Test for overall effect: $z = 0.20$ ($p = 0.84$)

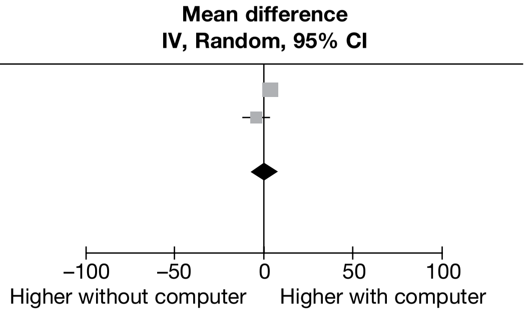


FIGURE 18 Meta-analysis SF-36 – bodily pain.

TABLE 16 Mean differences for computer – no computer

Outcome	Study	*n*	Mean diff.	SD (diff.)	95% CI	95% limits of agreement
Role emotional	Ryan 2002[124]	115	3.9	27.6	–1.1 to 8.9	–50.1 to 57.9
General health			1.0	10.5	–0.9 to 2.9	–19.6 to 21.5
Vitality			0.8	12.7	–1.6 to 3.1	–24.1 to 25.6
Bodily pain			0.6	16.4	–2.4 to 3.6	–31.6 to 32.8
Mental health			0.4	9.1	–1.3 to 2.0	–17.6 to 18.3
Role physical			0.2	14.7	–2.5 to 2.9	–28.6 to 29.0
Physical functioning			–0.5	16.4	–3.5 to 2.5	–32.6 to 31.6
Social functioning			–2.8	10.9	–4.8 to 0.8	–24.1 to 18.5

diff, difference.

between-group analysis, but other than social and physical functioning all results from this study go in the same direction as those from the between-group meta-analysis. It should be noted that the limits of agreement are very wide for all outcomes – this indicates that there could be considerable differences at an individual level. Although this may be of less concern to researchers, who are usually comparing groups, this would be much more of an issue if different modes were being used in clinical care and decisions on an individual basis.

Mode feature: administration and sensory stimuli

Seven between-subject studies compared modes in which there was a difference in administration.[104,112–114,116–118] All of these also had a difference in sensory stimuli, with auditory stimuli with interviewer administration and visual stimuli with self. Of these, five studies[104,112,114,116,118] provided data that could contribute to a meta-analysis. The results of the meta-analysis for each subscale of the SF-36 can be seen in *Figures 19–26* (forest plots in order of magnitude of pooled difference).

None of the scales show a significant difference between interviewer and self-administration, although all are in the direction of self-completion giving rise to higher scores. However, there was a high degree of heterogeneity between studies. The Jones *et al.* study[115] used the Veteran's SF-36, which was developed from the SF-36 to be specifically used in the Veteran's Health Administration.[127] Particular changes were made to the two subscales measuring role (physical and emotional) during the development process (see *Figures 22* and *25*). If the Jones *et al.* study[112] were to be excluded from the meta-analysis, the greatest impact would be on the effect for the 'role emotional' subscale, which would become significantly higher with interviewer administration [6.82 (95% CI 2.61 to 11.03)]; however, high levels of heterogeneity still remain. For the 'role physical' subscale, the effect changed sign, but was still not significant [2.24 (95% CI –3.28 to 7.76)]. For the other scales, three of the remaining six would also become positive, indicating higher scores for interviewer administration.

Two studies[113,118] provided data on differences in administration from the within-subject studies. The pooled estimators of effect can be seen in *Table 17*.

The pooled data from these two studies suggest higher scores for interviewer administration for all subscales, with all but bodily pain and vitality being significant. The impact on the two role subscales is in the order of 10 points; however, this is based on a total of only 250 patients.

	Interviewer			Self				
Study or subgroup	**Mean**	**SD**	**Total**	**Mean**	**SD**	**Total**	**Weight (%)**	**Mean difference IV, Random, 95% CI**
Bowling 1999[104]	89.6	19.3	2025	88.4	17.9	8801	21.6	1.20 (0.28 to 2.12)
Jones 2001[112]	36.86	27.84	1591	50.78	31.08	1659	21.3	–13.92 (–15.95 to –11.89)
Perkins 1998[114]	82.64	24.16	421	81.35	22.8	418	20.9	1.29 (–1.89 to 4.47)
Unruh 2003[116]	41.5	26.2	426	51.5	26.5	542	20.8	–10.00 (–13.34 to –6.66)
Weinberger 1996[118]	53.6	27.2	136	51.3	29.6	36	15.4	2.30 (–8.40 to 13.00)
Total (95% CI)			**4599**			**11,456**	**100.0**	**–4.17 (–11.98 to 3.64)**

Heterogeneity: $\tau^2 = 73.43$; $\chi^2 = 205.43$, df = 4 ($p < 0.00001$); $I^2 = 98\%$
Test for overall effect: $z = 1.05$ ($p = 0.30$)

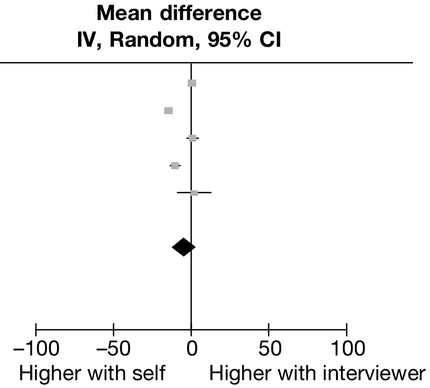


FIGURE 19 Meta-analysis SF-36 – physical functioning.

	Interviewer			Self				
Study or subgroup	**Mean**	**SD**	**Total**	**Mean**	**SD**	**Total**	**Weight (%)**	**Mean difference IV, Random, 95% CI**
Bowling 1999[104]	74	21.9	2017	73.5	19.9	8990	22.7	0.50 (–0.54 to 1.54)
Jones 2001[112]	33.18	21.84	1591	41.17	26.31	1659	22.2	–7.99 (–9.65 to –6.33)
Perkins 1998[114]	72.84	22.7	421	71.01	21.64	418	20.7	1.83 (–1.17 to 4.83)
Unruh 2003[116]	44.6	21.1	422	47.4	20.8	536	21.2	–2.80 (–5.47 to –0.13)
Weinberger 1996[118]	35.4	22	136	40.3	21.3	36	13.2	–4.90 (–12.78 to 2.98)
Total (95% CI)			**4587**			**11,639**	**100.0**	**–2.52 (–6.90 to 1.85)**

Heterogeneity: $\tau^2 = 21.70$; $\chi^2 = 78.94$, df = 4 ($p < 0.00001$); $I^2 = 95\%$
Test for overall effect: $z = 1.13$ ($p = 0.26$)

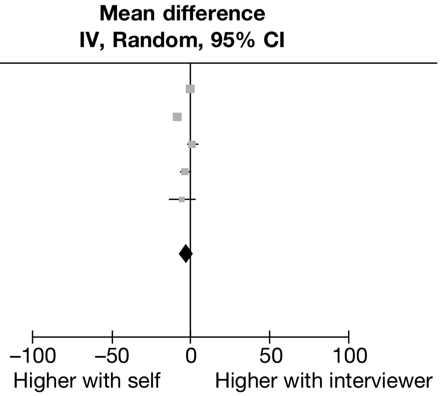


FIGURE 20 Meta-analysis SF-36 – general health perception.

	Interviewer			Self				
Study or subgroup	Mean	SD	Total	Mean	SD	Total	Weight (%)	Mean difference IV, Random, 95% CI
Bowling 1999[104]	89	20.8	2020	88	19.5	9124	22.1	1.00 (0.01 to 1.99)
Jones 2001[112]	43.37	30.92	1591	55.08	34.41	1659	21.6	–11.71 (–13.96 to –9.46)
Perkins 1998[114]	89.25	22.6	421	84.59	21.2	418	21.2	4.66 (1.70 to 7.62)
Unruh 2003[116]	70.8	29.5	425	70.9	25.5	550	20.8	–0.10 (–3.62 to –3.42)
Weinberger 1996[118]	62.2	29.4	136	69.8	26.6	36	14.2	–7.60 (–17.60 to 2.40)
Total (95% CI)			**4593**			**11,787**	**100.0**	**–2.42 (–8.68 to 3.83)**

Heterogeneity: $\tau^2 = 45.68$; $\chi^2 = 117.95$, df = 4 ($p < 0.00001$); $I^2 = 97\%$
Test for overall effect: $z = 0.76$ ($p = 0.45$)

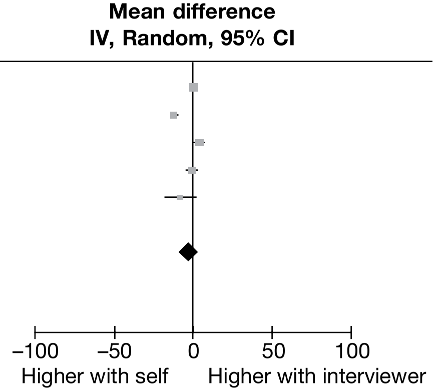


FIGURE 21 Meta-analysis SF-36 – social functioning.

	Interviewer			Self				
Study or subgroup	Mean	SD	Total	Mean	SD	Total	Weight (%)	Mean difference IV, Random, 95% CI
Bowling 1999[104]	84.2	32.7	2018	85.8	29.9	9058	22.1	–1.60 (–3.15 to –0.05)
Jones 2001[112]	18.92	30.96	1591	35.37	40.98	1659	21.9	–16.45 (–18.94 to –13.96)
Perkins 1998[114]	80.34	35.7	421	76.77	36.44	418	20.9	3.57 (–1.31 to 8.45)
Unruh 2003[116]	50.4	40.5	426	42.2	40.5	538	20.7	8.20 (3.05 to 13.35)
Weinberger 1996[118]	30.7	36	136	34.6	38.9	36	14.5	–3.90 (–17.97 to 10.17)
Total (95% CI)			**4592**			**11,709**	**100.0**	**–2.07 (–11.13 to 7.00)**

Heterogeneity: $\tau^2 = 96.26$; $\chi^2 = 135.65$, df = 4 ($p < 0.00001$); $I^2 = 97\%$
Test for overall effect: $z = 0.45$ ($p = 0.65$)

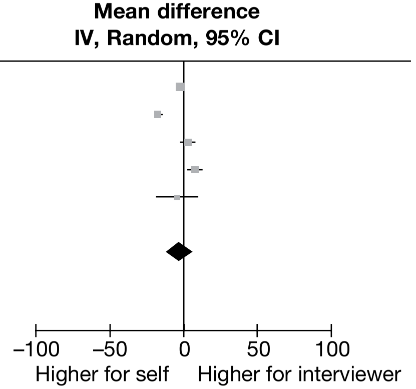


FIGURE 22 Meta-analysis SF-36 – role physical.

	Interviewer			Self				
Study or subgroup	Mean	SD	Total	Mean	SD	Total	Weight (%)	Mean difference IV, Random, 95% CI
Bowling 1999[104]	64.7	20.8	2018	61.1	19.6	8998	22.9	3.60 (2.61 to 4.59)
Jones 2001[112]	31.22	22.68	1591	35.4	26.12	1659	22.5	–4.18 (–5.88 to –2.50)
Perkins 1998[114]	63.22	22.78	421	60.57	21.48	418	21.0	2.65 (–0.35 to 5.65)
Unruh 2003[116]	47.2	24.7	425	51.6	19.3	545	21.2	–4.40 (–7.25 to –1.55)
Weinberger 1996[118]	35.3	24	136	43.2	24	36	12.3	–7.90 (–16.72 to 0.92)
Total (95% CI)			**4591**			**11,656**	**100.0**	**–1.46 (–5.97 to 3.05)**

Heterogeneity: $\tau^2 = 22.84$; $\chi^2 = 81.57$, df = 4 ($p < 0.00001$); $I^2 = 95\%$
Test for overall effect: $z = 0.64$ ($p = 0.53$)

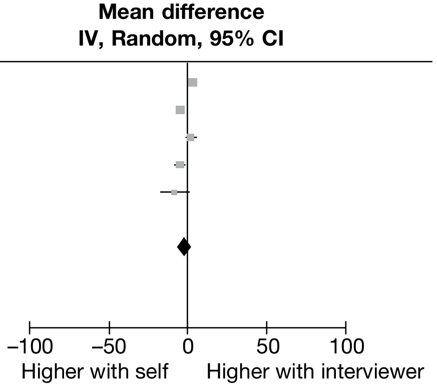


FIGURE 23 Meta-analysis SF-36 – vitality.

	Interviewer			Self				
Study or subgroup	Mean	SD	Total	Mean	SD	Total	Weight (%)	Mean difference IV, Random, 95% CI
Bowling 1999[104]	76.6	18.3	2019	73.8	17.2	8930	22.5	2.80 (1.93 to 3.67)
Jones 2001[112]	53.84	26.46	1591	58.76	28.13	1659	21.7	–4.92 (–6.80 to –3.04)
Perkins 1998[114]	80.38	17.42	421	76.33	16.28	418	21.2	4.05 (1.77 to 6.33)
Unruh 2003[116]	71.1	21.5	424	71.7	17.8	546	20.9	–0.60 (–3.13 to 1.93)
Weinberger 1996[118]	62.5	25.2	136	75	16.5	36	13.7	–12.50 (–19.35 to –5.65)
Total (95% CI)			**4591**			**11,589**	**100.0**	**–1.42 (–5.46 to 2.62)**

Heterogeneity: $\tau^2 = 18.64$; $\chi^2 = 76.77$, df = 4 ($p < 0.00001$); $I^2 = 95\%$
Test for overall effect: $z = 0.69$ ($p = 0.49$)

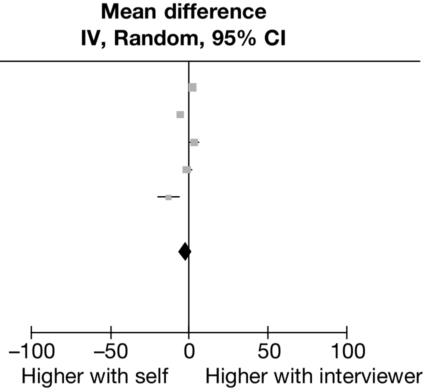


FIGURE 24 Meta-analysis SF-36 – mental health.

	Interviewer			Self				
Study or subgroup	Mean	SD	Total	Mean	SD	Total	Weight (%)	Mean difference IV, Random, 95% CI
Bowling 1999[104]	88	29.1	1919	82.9	31.8	8067	21.1	5.10 (3.62 to 6.58)
Jones 2001[112]	32.03	38.47	1591	54.48	43.91	1659	21.0	−22.45 (−25.29 to −19.61)
Perkins 1998[114]	88.76	27.99	421	80.36	33.47	418	20.8	8.40 (4.22 to 12.58)
Unruh 2003[116]	70.7	40.2	426	59.2	43.4	534	20.5	11.50 (6.20 to 16.80)
Weinberger 1996[118]	54.7	43.9	136	63.9	43.2	36	16.6	−9.20 (−25.12 to 6.72)
Total (95% CI)			**4493**			**10,714**	**100.0**	**−1.06 (−15.07 to 12.96)**

Heterogeneity: $\tau^2 = 241.71$; $\chi^2 = 320.74$, df = 4 ($p < 0.00001$); $I^2 = 99\%$
Test for overall effect: $z = 0.15$ ($p = 0.88$)

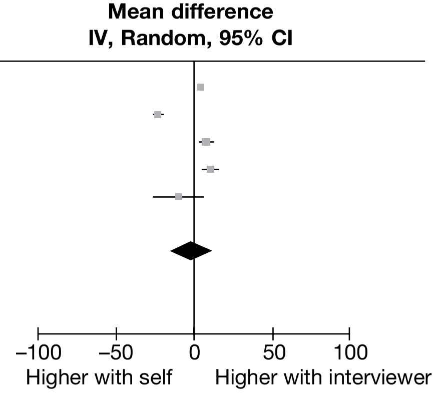


FIGURE 25 Meta-analysis SF-36 – role emotional.

	Interviewer			Self				
Study or subgroup	Mean	SD	Total	Mean	SD	Total	Weight (%)	Mean difference IV, Random, 95% CI
Bowling 1999[104]	82.5	24.8	2022	81.5	21.6	10105	23.9	1.00 (−0.16 to 2.16)
Jones 2001[112]	39.15	27.44	1591	45.77	29.88	1659	23.2	−6.62 (−8.59 to −4.65)
Perkins 1998[114]	77.75	27.64	421	73.73	24.41	421	21.0	4.02 (0.50 to 7.54)
Unruh 2003[116]	64.8	29.4	426	62.2	26.3	548	20.9	2.60 (−0.96 to 6.16)
Weinberger 1996[118]	43.3	27.3	136	46.7	26.4	36	10.9	−3.40 (−13.17 to 6.37)
Total (95% CI)			**4596**			**12,769**	**100.0**	**−0.28 (−4.63 to 4.07)**

Heterogeneity: $\tau^2 = 20.20$; $\chi^2 = 52.99$, df = 4 ($p < 0.00001$); $I^2 = 92\%$
Test for overall effect: $z = 0.13$ ($p = 0.90$)

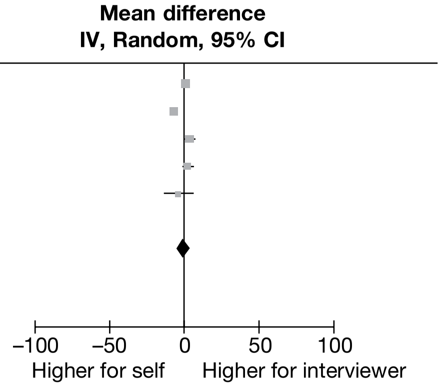


FIGURE 26 Meta-analysis SF-36 – bodily pain.

Mode feature: telephone

Three between-subject studies[112,114,118] provided data for consideration of the impact of the telephone mode feature. The results of the meta-analysis for each subscale of the SF-36 can be seen in *Figures 27–34* (forest plots by order of magnitude of pooled difference).

All of the subscales had differences in the direction of giving higher scores without a telephone. As for the previous analysis, there were high levels of heterogeneity. The two largest effects were for the 'role physical' and 'role emotional' scales (see *Figures 30* and *31*). Excluding the Jones *et al.* study[112] from these (as it was using Veteran's SF-36) would considerably reduce the estimated mean difference (to –0.77 and 1.45, respectively).

Two studies[118,123] provided data on differences in telephone administration from the within-subject studies. The pooled estimators of effect can be seen in *Table 18*.

All of the subscales except 'vitality' show a mean difference in the direction of higher scores without a telephone, which is consistent with the results from the between-group studies. Only 'role physical' shows a significant difference.

Meta-analysis of the Minnesota Multiphasic Personality Inventory

The second most frequently occurring measure from the studies included was the MMPI.[128] The MMPI was developed in the 1930s at Minnesota University as a comprehensive personality test that could be used to detect psychiatric problems. The MMPI consists of 14 scaled scores. Ten clinical scales are included to indicate different psychiatric conditions (hypochondriasis,

TABLE 17 Mean differences for interviewer – self

Outcome	*n*	Mean diff.	SD (diff.)	95% CI	95% limits of agreement
Role emotional	250	12.8	41.6	7.6 to 17.9	–68.7 to 94.2
Role physical	250	10.0	31.8	5.8 to 14.2	–56.5 to 76.5
Social functioning	250	5.5	22.3	2.7 to 8.2	–38.3 to 49.3
Physical functioning	250	5.2	18.1	3.0 to 7.5	–30.3 to 40.7
General health	250	3.5	14.0	1.7 to 5.2	–24.0 to 30.9
Mental health	250	2.9	15.0	1.0 to 4.7	–26.5 to 32.2
Bodily pain	250	1.5	21.2	–1.1 to 4.2	–40.0 to 43.1
Vitality	250	0.7	16.9	–1.4 to 2.8	–32.3 to 33.8

diff., difference.

TABLE 18 Mean differences for telephone – no telephone

Outcome	*n*	Mean diff	SD (diff.)	95% CI	95% limits of agreement
Role physical	73	–6.1	22.6	–11.3 to –0.9	–50.4 to 38.2
Social functioning	73	–4.1	25.9	–10.0 to 1.9	–54.8 to 46.7
Role emotional	73	–3.9	25.9	–9.8 to 2.0	–54.6 to 46.8
Physical functioning	73	–2.1	12.6	–5.0 to 0.8	–26.7 to 22.5
Bodily pain	73	–0.7	15.1	–4.1 to 2.8	–30.3 to 29.0
General health	73	–0.3	13.6	–3.4 to 2.8	–27.0 to 26.4
Mental health	73	–0.2	10.6	–2.6 to 2.2	–20.9 to 20.6
Vitality	73	0.8	14.9	–2.7 to 4.2	–28.4 to 29.9

	Telephone			No telephone				Mean difference	Mean difference
Study or subgroup	Mean	SD	Total	Mean	SD	Total	Weight (%)	IV, Random, 95% CI	IV, Random, 95% CI
Jones 2001[112]	18.92	30.96	1591	35.37	40.98	1659	35.6	16.45 (–18.94 to –13.96)	
Perkins 1998[114]	80.34	35.7	421	76.77	36.44	418	34.6	3.57 (–1.31 to 8.45)	
Weinberger 1996[118]	23.9	32.5	47	34.4	38	125	29.7	–10.50 (–21.93 to 0.93)	
Total (95% CI)			**2059**			**2202**	**100.0**	**–7.74 (–22.74 to 7.25)**	

Heterogeneity: $\tau^2 = 162.80$; $\chi^2 = 51.35$, df = 2 ($p < 0.00001$); $I^2 = 96\%$
Test for overall effect: $z = 1.01$ ($p = 0.31$)

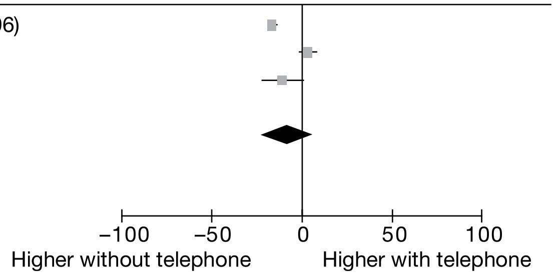


FIGURE 27 Meta-analysis SF-36 – role physical.

	Telephone			No telephone				Mean difference	Mean difference
Study or subgroup	Mean	SD	Total	Mean	SD	Total	Weight (%)	IV, Random, 95% CI	IV, Random, 95% CI
Jones 2001[112]	32.03	38.47	1591	54.48	43.91	1659	34.7	–22.45 (–25.29 to –19.61)	
Perkins 1998[114]	88.76	27.99	421	80.36	33.47	418	34.5	8.40 (4.22 to 12.58)	
Weinberger 1996[118]	50.4	43.3	47	58.9	43.9	125	30.9	–8.50 (–23.08 to 6.08)	
Total (95% CI)			**2059**			**2202**	**100.0**	**–7.51 (–31.52 to 16.50)**	

Heterogeneity: $\tau^2 = 430.78$; $\chi^2 = 143.79$, df = 2 ($p < 0.00001$); $I^2 = 99\%$
Test for overall effect: $z = 0.61$ ($p = 0.54$)

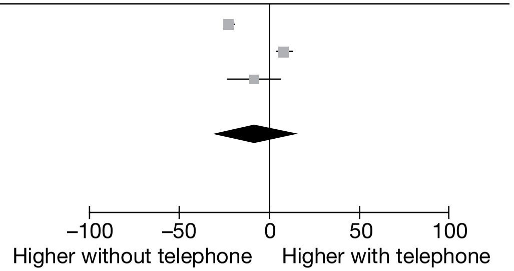


FIGURE 28 Meta-analysis SF-36 – role emotional.

	Telephone			No telephone				
Study or subgroup	Mean	SD	Total	Mean	SD	Total	Weight (%)	Mean difference IV, Random, 95% CI
Jones 2001[112]	43.37	30.92	1591	55.08	34.41	1659	35.3	–11.71 (–13.96 to –9.46)
Perkins 1998[114]	89.25	22.6	421	84.59	21.2	418	35.0	4.66 (1.70 to 7.62)
Weinberger 1996[118]	54.5	29.2	47	67.3	28.7	125	29.6	–12.80 (–22.55 to –3.05)
Total (95% CI)			**2059**			**2202**	**100.0**	**–6.30 (–19.17 to 6.58)**

Heterogeneity: $\tau^2 = 120.82$; $\chi^2 = 76.33$, df = 2 ($p < 0.00001$); $I^2 = 97\%$
Test for overall effect: $z = 0.96$ ($p = 0.34$)

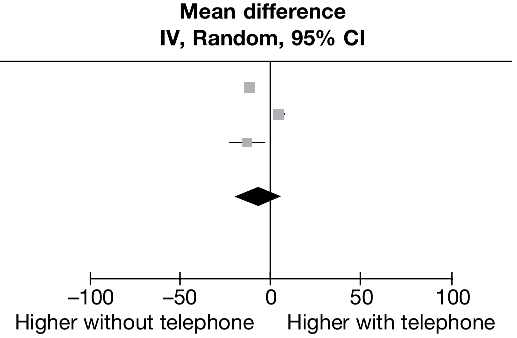


FIGURE 29 Meta-analysis SF-36 – social functioning.

	Telephone			No telephone				
Study or subgroup	Mean	SD	Total	Mean	SD	Total	Weight (%)	Mean difference IV, Random, 95% CI
Jones 2001[112]	36.86	27.84	1591	50.78	31.08	1659	35.6	–13.92 (–13.95 to –11.89)
Perkins 1998[114]	82.64	24.16	421	81.35	22.8	418	35.1	1.29 (–1.89 to 4.47)
Weinberger 1996[118]	49	27.9	47	54.7	27.7	125	29.3	–5.70 (–15.04 to 3.64)
Total (95% CI)			**2059**			**2202**	**100.0**	**–6.18 (–17.96 to 5.61)**

Heterogeneity: $\tau^2 = 100.53$; $\chi^2 = 63.16$, df = 2 ($p < 0.00001$); $I^2 = 97\%$
Test for overall effect: $z = 1.03$ ($p = 0.30$)

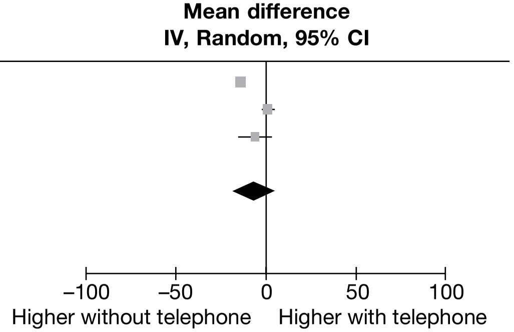


FIGURE 30 Meta-analysis SF-36 – physical functioning.

	Telephone			No telephone				
Study or subgroup	Mean	SD	Total	Mean	SD	Total	Weight (%)	Mean difference IV, Random, 95% CI
Jones 2001[112]	33.18	21.84	1591	41.17	26.31	1659	37.0	−7.99 (−9.65 to −6.33)
Perkins 1998[114]	72.84	22.7	421	71.01	21.64	418	35.5	1.83 (−1.17 to 4.83)
Weinberger 1996[118]	31.9	22.4	47	38.1	21.6	125	27.5	−6.20 (−13.64 to 1.24)
Total (95% CI)			**2059**			**2202**	**100.0**	**−4.01 (−11.49 to 3.47)**

Heterogeneity: $\tau^2 = 38.68$; $\chi^2 = 31.52$, df = 2 ($p < 0.00001$); $I^2 = 94\%$
Test for overall effect: $z = 1.05$ ($p = 0.29$)

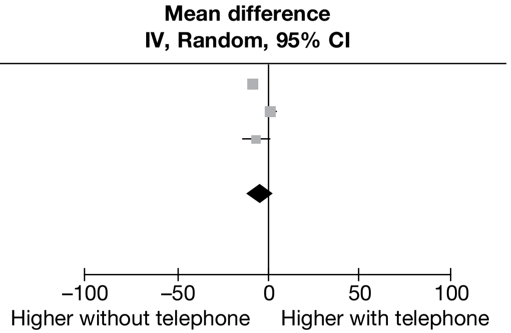


FIGURE 31 Meta-analysis SF-36 – general health perception.

	Telephone			No telephone				
Study or subgroup	Mean	SD	Total	Mean	SD	Total	Weight (%)	Mean difference IV, Random, 95% CI
Jones 2001[112]	31.22	22.68	1591	35.4	26.12	1659	39.0	−4.18 (−5.86 to −2.50)
Perkins 1998[114]	63.22	22.78	421	60.57	21.48	418	36.5	2.65 (−0.35 to 5.65)
Weinberger 1996[118]	29.6	21.1	47	39.7	25	125	24.5	−10.10 (−17.56 to −2.64)
Total (95% CI)			**2059**			**2202**	**100.0**	**−3.14 (−9.02 to 2.75)**

Heterogeneity: $\tau^2 = 22.35$; $\chi^2 = 18.99$, df = 2 ($p < 0.0001$); $I^2 = 89\%$
Test for overall effect: $z = 1.04$ ($p = 0.30$)

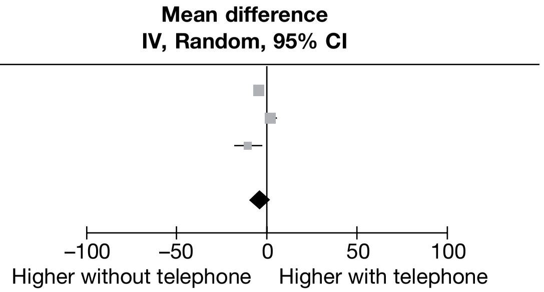


FIGURE 32 Meta-analysis SF-36 – vitality.

	Telephone			No telephone				
Study or subgroup	Mean	SD	Total	Mean	SD	Total	Weight (%)	Mean difference IV, Random, 95% CI
Jones 2001[112]	53.84	26.46	1591	58.76	28.13	1659	37.7	–4.92 (–6.80 to –3.04)
Perkins 1998[114]	80.38	17.42	421	76.33	16.28	418	37.3	4.05 (1.77 to 6.33)
Weinberger 1996[118]	58.9	28.3	47	67.5	21.7	125	24.9	–8.60 (–17.54 to 0.34)
Total (95% CI)			**2059**			**2202**	**100.0**	**–2.49 (–9.98 to 5.00)**

Heterogeneity: $\tau^2 = 37.74$; $\chi^2 = 37.91$, df = 2 ($p < 0.00001$); $I^2 = 95\%$
Test for overall effect: $z = 0.65$ ($p = 0.51$)

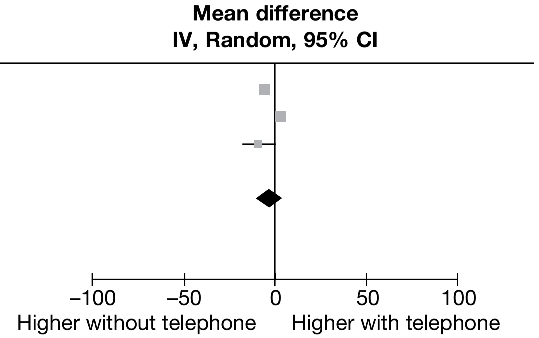


FIGURE 33 Meta-analysis SF-36 – mental health.

	Telephone			No telephone				
Study or subgroup	Mean	SD	Total	Mean	SD	Total	Weight (%)	Mean difference IV, Random, 95% CI
Jones 2001[112]	39.15	27.44	1591	45.77	29.88	1659	37.5	–6.62 (–8.59 to –4.65)
Perkins 1998[114]	77.75	27.64	421	73.73	24.41	418	35.8	4.02 (0.49 to 7.55)
Weinberger 1996[118]	40.1	25.1	47	45.5	27.8	125	26.7	–5.40 (–14.07 to 3.27)
Total (95% CI)			**2059**			**2202**	**100.0**	**–2.49 (–10.62 to 5.64)**

Heterogeneity: $\tau^2 = 44.86$; $\chi^2 = 26.71$, df = 2 ($p < 0.00001$); $I^2 = 93\%$
Test for overall effect: $z = 0.60$ ($p = 0.55$)

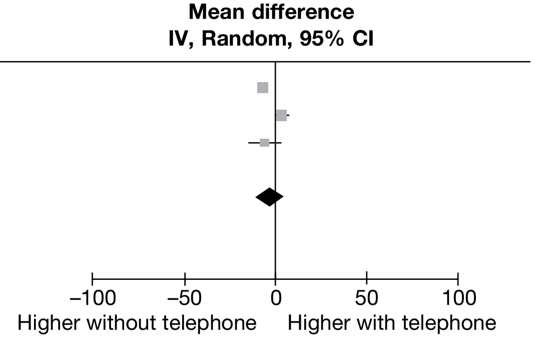


FIGURE 34 Meta-analysis SF-36 – bodily pain.

depression, hysteria, psychopathic deviation, masculinity–femininity, paranoia, psychasthenia, schizophrenia, hypomania and social introversion). The four remaining scales are included to safeguard against participants giving false results. The four validity scales are 'cannot say', lie, infrequency and defensiveness. The raw scores from each scale are hard to understand and are therefore transformed into standardised version of the score (*T*-score). Each standardised scale is scored on a range from 0 to 100 to aid interpretation.

Nine studies[129–137] published between 1994 and 2003 used the MMPI. Not all of the studies reported all subscales. As for the SF-36, the impact of different modes of using the MMPI is assessed using weighted pooled measures of agreement for within-subject comparisons and random-effects meta-analysis for between-subject comparisons. Unlike the SF-36, however, the only mode comparison available for the MMPI was 'computer' versus 'not computer'.

Of the nine studies[129–137] that reported the use of the MMPI, data were available from three studies[129,131,134] for between-subject comparisons only, one study[136] for within-subject comparisons only and three studies[132,133,135] for both. *Table 19* summaries the information available from each study.

TABLE 19 Included studies using MMPI

Paper	Country	Population	Design	Comp	Adm	Tel	Sens	*Data availability*
Between-subject comparisons								
Biskin (1977)[129]	USA	Psychology students	Randomised trial	y	n	n	y	Data available
Evan (1969)[130]	USA	Psychology students	Randomised trial	y	n	n	y	No data
Hart (1985)[131]	USA	Male psychiatric referrals	Randomised trial	y	n	n	y	Data available for all scales other than the psychopathic deviant scale
Honaker (1988)[132]	USA	General population	Repeated measures	y	n	n	y	Data taken prior to crossover
Lambert (1987)[133]	USA	Substance abusers	Latin squares	y	n	n	y	Data taken prior to crossover
Locke (1995)[134]	USA	Psychology students	Randomised trial	y	n	n	y	Data available only for the F scale[a]
White (1985)[135]	USA	Psychology students	Crossover	y	n	n	y	Data taken prior to crossover
Within-subject comparisons								
Honaker (1988)[132]	USA	General population	Repeated measures	y	n	n	y	Data taken combining order groups
Lambert (1987)[133]	USA	Substance abusers	Latin squares	y	n	n	y	Data taken combining order groups
Pinsoneault (1996)[136]	USA	Psychology students	Randomised crossover	y	n	n	y	Data taken combining order groups
Shuldberg (1988)[137]	USA	Psychology students	Crossover	y	n	n	y	No SDs
White (1985)[135]	USA	Psychology students	Crossover	y	n	n	y	Data taken combining order groups

Adm, administration; Comp, computerisation; n, no; Sens, sensory stimuli; Tel, telephone; y, yes.
[a] See *Appendix 7* for details of F scale.
Shaded cells indicate that these studies contributed to the analysis of that mode feature.

Between-subject comparisons

Six studies had data available that could contribute to the meta-analysis. Data taken from studies with a within-subject design have used data prior to any crossover. One study[134] only had data available for one of the 14 scales. One study[131] had data available for all scales other than 'psychopathic deviant'. One study[130] had no data that could be used for the meta-analysis.

Within-subject comparisons

Four studies had data available that could contribute to the within-subjects meta-analysis. One study, Lambert *et al.*[133] had data available for all subscales other than the 'cannot say' scale. One study Shuldberg[137] provided only mean scores, so could not contribute to the meta-analysis.

Mode feature: computer

All included studies that measured MMPI did so comparing 'computer administered' versus 'not computer administered'. The results of the meta-analysis for each subscale of the MMPI can be viewed in *Figures 35–48* (forest plots in order of magnitude of difference).

The 'cannot say' scale of the MMPI suggested higher scores when administered without a computer, with a mean difference of over seven points on the scale (see *Figure 35*). Although this could imply that participants view a computer terminal as a more private mode of data capture, and are, hence, less likely to leave a question blank than if they had to complete the MMPI with another form of data capture. It is more likely that the computer-completed measures did not allow for leaving items unanswered without justification. None of the clinical scales showed any significant differences (see *Figures 39–48*).

The combined results for the within-subjects studies are given in *Table 20*.

There were no significant differences between modes of administration for any of the subscales in the within-subject studies.

TABLE 20 Mean differences for computer – no computer

Outcome	*n*	Mean difference	SD (diff.)	95% CI	95% limits of agreement
Hypochondriasis	172	–0.7	5.2	–1.4 to 0.1	–10.9 to 9.5
Depression	172	–0.5	6.2	–1.4 to 0.5	–12.6 to 11.7
Hysteria	172	–0.5	5.7	–1.3 to 0.4	–11.6 to 10.6
Psychopathic deviation	172	–0.2	5.1	–1.0 to 0.6	–10.1 to 9.7
Masculinity–femininity	172	0.0	5.2	–0.7 to 0.8	–10.1 to 10.1
Paranoia	172	–0.7	4.9	–1.5 to 0.0	–10.3 to 8.9
Psychasthenia	172	–0.7	7.4	–1.8 to 0.4	–15.3 to 13.9
Schizophrenia	172	–0.8	10.0	–2.3 to 0.7	–20.4 to 18.7
Hypomania	172	–0.4	4.0	–1.0 to 0.2	–8.3 to 7.5
Social introversion	172	–0.5	6.5	–1.5 to 0.4	–13.3 to 12.2
Cannot say	97	–0.1	0.5	–0.2 to 0.0	–1.1 to 0.9
L	172	0.1	1.7	–0.2 to 0.4	–3.2 to 3.4
F	172	–0.5	5.9	–1.4 to 0.4	–12.1 to 11.0
K	172	–0.3	3.9	–0.9 to 0.3	–8.0 to 7.4

diff., difference.

Study or subgroup	Computer Mean	Computer SD	Computer Total	No computer Mean	No computer SD	No computer Total	Weight (%)	Mean difference IV, Random, 95% CI
Biskin 1977[129]	15.757	18.435	37	1.889	3.002	45	23.3	13.87 (7.86 to 19.87)
White 1986[135]	15.16	20.357	50	1.5	5.19	50	23.5	13.66 (7.84 to 19.48)
Honaker 1989[132]	1.4	3.099	40	1.1	3.111	40	27.9	0.30 (1.06 to 1.66)
Lambert 1987[133]	0	0	38	1.82	6.16	37		Not estimable
Hart 1985[131]	4.4	6.7	10	1.6	2.1	10	25.4	2.80 (–1.55 to 7.15)
Total (95% CI)			**137**			**145**	**100.0**	**7.23 (0.29 to 14.17)**

Heterogeneity: $\tau^2 = 44.53$; $\chi^2 = 36.29$, df = 3 ($p < 0.00001$); $I^2 = 92\%$

Test for overall effect: $z = 2.04$ ($p = 0.04$)

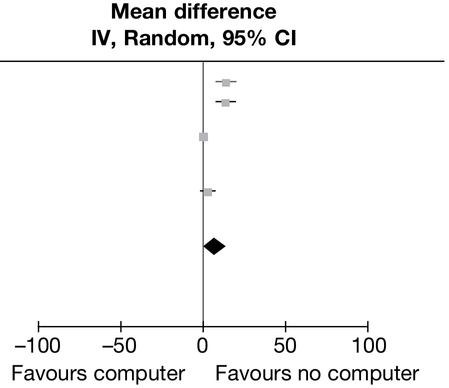


FIGURE 35 Meta-analysis MMPI – cannot say.

Study or subgroup	Computer Mean	Computer SD	Computer Total	No computer Mean	No computer SD	No computer Total	Weight (%)	Mean difference IV, Random, 95% CI
Biskin 1977[129]	6.432	4.616	37	6.113	3.817	45	9.3	0.32 (–1.54 to 2.18)
White 1986[135]	5.02	3.325	50	5.44	3.162	50	19.8	–0.42 (–1.69 to 0.85)
Honaker 1989[132]	55.9	8.415	40	54.65	7.106	40	2.8	1.25 (–2.16 to 4.66)
Lambert 1987[133]	12.19	7.91	38	11.45	8.54	37	2.3	0.74 (–2.99 to 4.47)
Hart 1985[131]	63.8	13.7	10	67.9	13.5	10	0.2	–4.10 (–16.02 to 7.82)
Locke 1995[134]	4	2.1	54	3.8	2.22	108	65.6	0.20 (–0.50 to 0.90)
Total (95% CI)			**229**			**290**	**100.0**	**0.12 (–0.45 to 0.69)**

Heterogeneity: $\tau^2 = 0.00$; $\chi^2 = 1.80$, df = 5 ($p = 0.88$); $I^2 = 0\%$

Test for overall effect: $z = 0.41$ ($p = 0.68$)

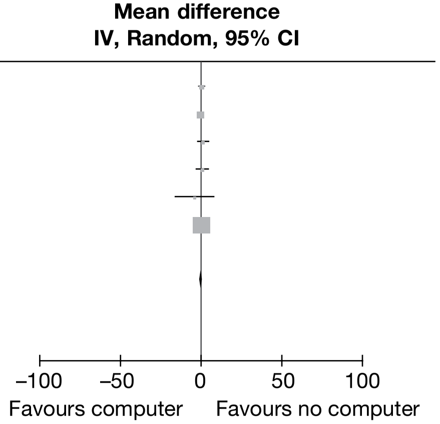


FIGURE 36 Meta-analysis MMPI – F.

	Computer			No computer				
Study or subgroup	Mean	SD	Total	Mean	SD	Total	Weight (%)	Mean difference IV, Random, 95% CI
Biskin 1977[129]	13.297	5.72	37	14.311	5.134	45	19.7	–1.01 (–3.39 to 1.36)
White 1986[135]	13.22	3.846	50	13.28	4.43	50	42.1	–0.06 (–1.69 to 1.57)
Honaker 1989[132]	54.15	9.253	40	54.6	8.857	40	7.1	–0.45 (–4.42 to 3.52)
Lambert 1987[133]	10.35	5	38	12.16	3.43	37	29.7	–1.81 (–3.75 to 0.13)
Hart 1985[131]	50.1	9.8	10	49.1	9.8	10	1.5	1.00 (–7.59 to 9.59)
Total (95% CI)			**175**			**182**	**100.0**	**–0.78 (–1.83 to 0.28)**

Heterogeneity: $\tau^2 = 0.00$; $\chi^2 = 2.07$, df = 4 ($p = 0.72$); $I^2 = 0\%$
Test for overall effect: $z = 1.45$ ($p = 0.15$)

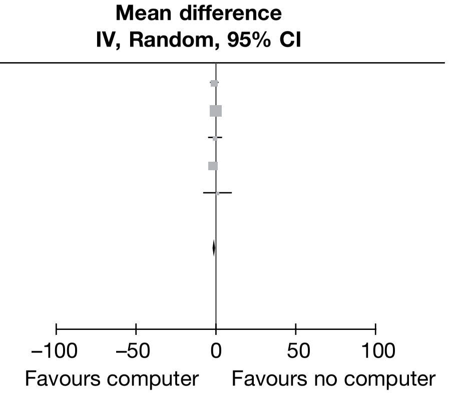


FIGURE 37 Meta-analysis MMPI – K.

	Computer			No computer				
Study or subgroup	Mean	SD	Total	Mean	SD	Total	Weight (%)	Mean difference IV, Random, 95% CI
Biskin 1977[129]	2.72	2.232	37	2.622	2.124	45	28.4	0.10 (–0.85 to 1.05)
White 1986[135]	2.82	1.706	50	2.22	1.598	50	42.4	0.60 (–0.05 to 1.25)
Honaker 1989[132]	48	7.845	40	45.85	6.107	40	4.1	2.15 (–0.93 to 5.23)
Lambert 1987[133]	3.49	2.46	38	4.1	2.28	37	24.3	–0.61 (–1.68 to 0.46)
Hart 1985[131]	49.6	5.7	10	51.6	9.5	10	0.9	–2.00 (–8.87 to 4.87)
Total (95% CI)			**175**			**182**	**100.0**	**0.21 (- 0.44 to 0.85)**

Heterogeneity: $t^2 = 0.15$; $\chi^2 = 5.56$, df = 4 ($p = 0.23$); $I^2 = 28\%$
Test for overall effect: $z = 0.62$ ($p = 0.53$)

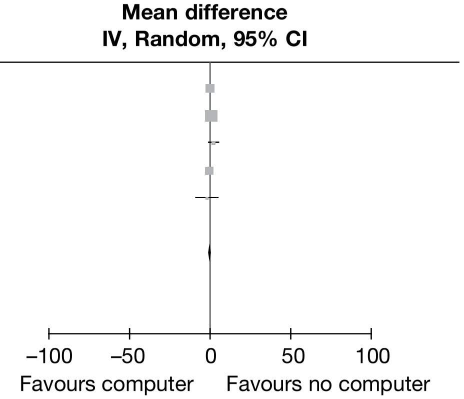


FIGURE 38 Meta-analysis MMPI – L.

	Computer			No computer				
Study or subgroup	**Mean**	**SD**	**Total**	**Mean**	**SD**	**Total**	**Weight (%)**	**Mean difference IV, Random, 95% CI**
Biskin 1977[129]	28.892	10.661	37	26.4	10.532	45	19.7	2.49 (–2.12 to 7.10)
White 1986[135]	25.28	8.519	50	23.3	9.471	50	33.5	1.98 (–1.55 to 5.51)
Honaker 1989[132]	48.45	7.732	40	48.1	10.042	40	27.1	0.35 (–3.58 to 4.28)
Lambert 1987[133]	33.19	11.88	38	33.47	11.13	37	15.4	–0.28 (–5.49 to 4.93)
Hart 1985[131]	59.5	11	10	62.7	11.5	10	4.3	–3.20 (–13.06 to 6.66)
Total (95% CI)			**175**			**182**	**100.0**	**1.07 (–0.98 to 3.11)**

Heterogeneity: $\tau^2 = 0.00$; $\chi^2 = 1.73$, df = 4 ($p = 0.79$); $I^2 = 0\%$
Test for overall effect: $z = 1.02$ ($p = 0.31$)

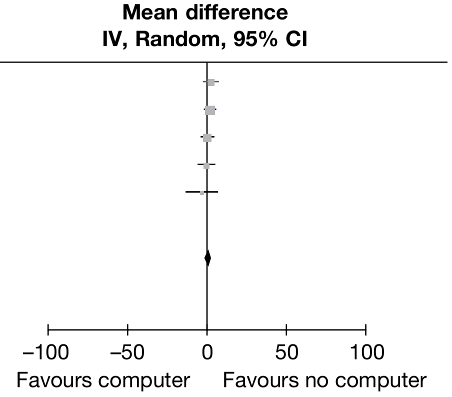


FIGURE 39 Meta-analysis MMPI – social introversion.

	Computer			No computer				
Study or subgroup	**Mean**	**SD**	**Total**	**Mean**	**SD**	**Total**	**Weight (%)**	**Mean difference IV, Random, 95% CI**
Biskin 1977[129]	28.541	5.714	37	30.311	5.368	45	26.8	–1.77 (–4.19 to 0.65)
White 1986[135]	32.72	8.261	50	30.18	7.041	50	23.2	2.54 (–0.47 to 5.55)
Honaker 1989[132]	54.85	13.653	40	53.9	12.055	40	11.9	0.95 (–4.69 to 6.59)
Lambert 1987[133]	23.95	4.53	38	25.18	4.8	37	28.8	–1.23 (–3.34 to 0.88)
Hart 1985[131]	62.1	8.8	10	69.7	6.3	10	9.3	–7.60 (–14.31 to –0.89)
Total (95% CI)			**175**			**182**	**100.0**	**–0.83 (–3.19 to 1.53)**

Heterogeneity: $\tau^2 = 3.87$; $\chi^2 = 9.83$, df = 4 ($p = 0.04$); $I^2 = 59\%$
Test for overall effect: $z = 0.69$ ($p = 0.49$)

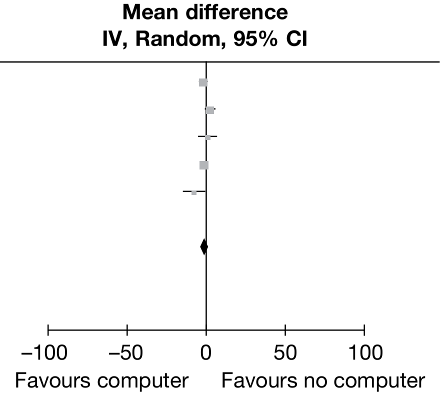


FIGURE 40 Meta-analysis MMPI – masculinity–femininity.

	Computer			No computer				
Study or subgroup	Mean	SD	Total	Mean	SD	Total	Weight (%)	Mean difference IV, Random, 95% CI
Biskin 1977[129]	18.243	5.118	37	19.311	5.888	45	27.6	–1.07 (–3.45 to 1.32)
White 1986[135]	18.32	3.759	50	18.84	5.158	50	50.1	–0.52 (–2.29 to 1.25)
Honaker 1989[132]	54.95	10.602	40	54.6	9.299	40	8.2	0.35 (–4.02 to 4.72)
Lambert 1987[133]	27.78	8.43	38	28.55	6.62	37	13.4	–0.77 (–4.20 to 2.66)
Hart 1985[131]	73.9	20.7	10	84.9	12.5	10	0.7	–11.00 (–25.99 to 3.99)
Total (95% CI)			**175**			**182**	**100.0**	**–0.71 (–1.96 to 0.55)**

Heterogeneity: $\tau^2 = 0.00$; $\chi^2 = 2.17$, df = 4 ($p = 0.70$); $I^2 = 0\%$
Test for overall effect: $z = 1.11$ ($p = 0.27$)

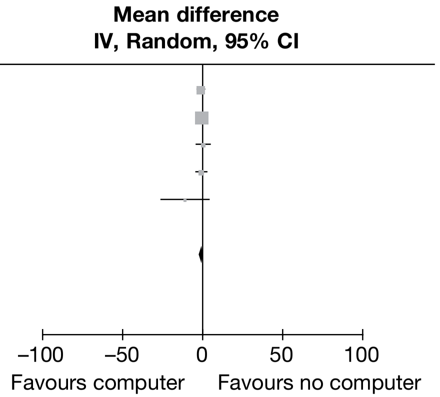


FIGURE 41 Meta-analysis MMPI – depression.

	Computer			No computer				
Study or subgroup	Mean	SD	Total	Mean	SD	Total	Weight (%)	Mean difference IV, Random, 95% CI
Biskin 1977[129]	9.432	2.93	37	11.067	3.434	45	34.2	–1.63 (–3.01 to –0.26)
White 1986[135]	9.64	2.85	50	9.6	2.999	50	40.8	0.04 (–1.11 to 1.19)
Honaker 1989[132]	56.7	10.493	40	55.65	8.25	40	6.5	1.05 (–3.09 to 5.19)
Lambert 1987[133]	13	5.05	38	13.9	5.11	37	17.5	–0.90 (–3.20 to 1.40)
Hart 1985[131]	61.8	12.1	10	69.4	12.6	10	1.0	–7.60 (–18.52 to 3.32)
Total (95% CI)			**175**			**182**	**100.0**	**–0.71 (–1.81 to 0.40)**

Heterogeneity: $\tau^2 = 0.43$; $\chi^2 = 5.61$, df = 4 ($p = 0.23$); $I^2 = 29\%$
Test for overall effect: $z = 1.26$ ($p = 0.21$)

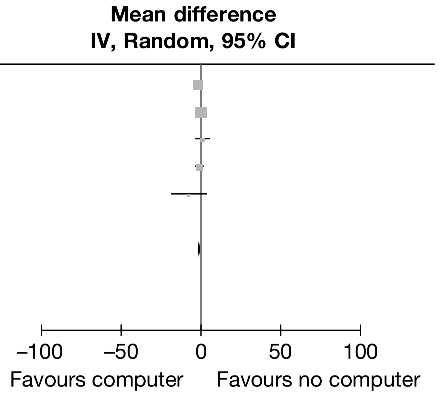


FIGURE 42 Meta-analysis MMPI – paranoia.

Study or subgroup	Computer Mean	Computer SD	Computer Total	No computer Mean	No computer SD	No computer Total	Weight (%)	Mean difference IV, Random, 95% CI
Biskin 1977[129]	18.487	5.026	37	19.289	4.294	45	29.1	–0.80 (–2.85 to 1.25)
White 1986[135]	18.88	4.237	50	18.74	4.7	50	39.6	0.14 (–1.61 to 1.89)
Honaker 1989[132]	63.9	12.426	40	63.4	11.302	40	4.5	0.50 (–4.71 to 5.71)
Lambert 1987[133]	21.35	5.28	38	22.18	4.3	37	25.7	–0.83 (–3.01 to 1.35)
Hart 1985[131]	64.5	8.5	10	69.5	15.1	10	1.1	–5.00 (–15.74 to 5.74)
Total (95% CI)			**175**			**182**	**100.0**	**–0.42 (–1.53 to 0.68)**

Heterogeneity: $\tau^2 = 0.00$; $\chi^2 = 1.48$, df = 4 ($p = 0.83$); $I^2 = 0\%$

Test for overall effect: $z = 0.75$ ($p = 0.45$)

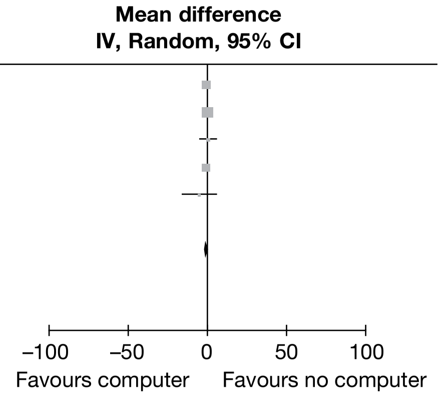


FIGURE 43 Meta-analysis MMPI – hypomania.

Study or subgroup	Computer Mean	Computer SD	Computer Total	No computer Mean	No computer SD	No computer Total	Weight (%)	Mean difference IV, Random, 95% CI
Biskin 1977[129]	5.054	3.274	37	5.422	4.218	45	31.8	–0.37 (–1.99 to 1.25)
White 1986[135]	6.48	2.901	50	5.86	3.567	50	51.5	0.62 (–0.65 to 1.89)
Honaker 1989[132]	53.15	7.303	40	51.75	7.301	40	8.2	1.40 (–1.80 to 4.60)
Lambert 1987[133]	12.51	6.96	38	13.24	7.12	37	8.2	–0.73 (–3.92 to 2.46)
Hart 1985[131]	67.4	19.5	10	72.2	14.8	10	0.4	–4.80 (–19.97 to 10.37)
Total (95% CI)			**175**			**182**	**100.0**	**0.24 (–0.68 to 1.15)**

Heterogeneity: $\tau^2 = 0.00$; $\chi^2 = 2.17$, df = 4 ($p = 0.71$); $I^2 = 0\%$

Test for overall effect: $z = 0.51$ ($p = 0.61$)

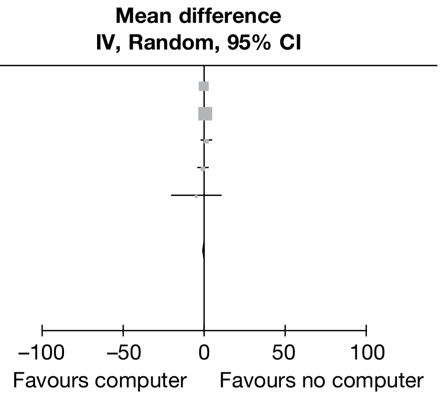


FIGURE 44 Meta-analysis MMPI – hypochondriasis.

	Computer			No computer				Mean difference
Study or subgroup	Mean	SD	Total	Mean	SD	Total	Weight (%)	IV, Random, 95% CI
Biskin 1977[129]	18.865	3.772	37	21	4.39	45	28.0	–2.14 (–3.90 to –0.37)
White 1986[135]	19.56	4.403	50	20.02	4.037	50	28.7	–0.46 (–2.12 to 1.20)
Honaker 1989[132]	59.35	8.106	40	55.05	6.266	40	19.8	4.30 (1.12 to 7.48)
Lambert 1987[133]	24.38	6.95	38	25.68	6.64	37	20.3	–1.30 (–4.38 to 1.78)
Hart 1985[131]	68.3	15.7	10	70.3	9.9	10	3.2	–2.00 (–13.50 to 9.50)
Total (95% CI)			**175**			**182**	**100.0**	**–0.21 (–2.37 to 1.96)**

Heterogeneity: $\tau^2 = 3.54$; $\chi^2 = 12.32$, df = 4 ($p = 0.02$); $I^2 = 68\%$

Test for overall effect: $z = 0.19$ ($p = 0.85$)

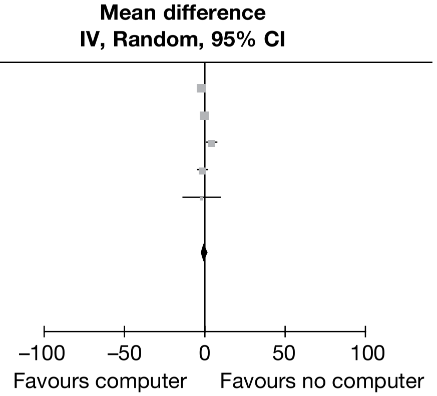


FIGURE 45 Meta-analysis MMPI – hysteria.

	Computer			No computer				Mean difference
Study or subgroup	Mean	SD	Total	Mean	SD	Total	Weight (%)	IV, Random, 95% CI
Biskin 1977[129]	18.865	3.772	37	21	4.39	45	28.0	–2.14 (–3.90 to –0.37)
White 1986[135]	19.56	4.403	50	20.02	4.037	50	28.7	–0.46 (–2.12 to 1.20)
Honaker 1989[132]	59.35	8.106	40	55.05	6.266	40	19.8	4.30 (1.12 to 7.48)
Lambert 1987[133]	24.38	6.95	38	25.68	6.64	37	20.3	–1.30 (–4.38 to 1.78)
Hart 1985[131]	68.3	15.7	10	70.3	9.9	10	3.2	–2.00 (–13.50 to 9.50)
Total (95% CI)			**175**			**182**	**100.0**	**–0.21 (–2.37 to 1.96)**

Heterogeneity: $\tau^2 = 3.54$; $\chi^2 = 12.32$, df = 4 ($p = 0.02$); $I^2 = 68\%$

Test for overall effect: $z = 0.19$ ($p = 0.85$)

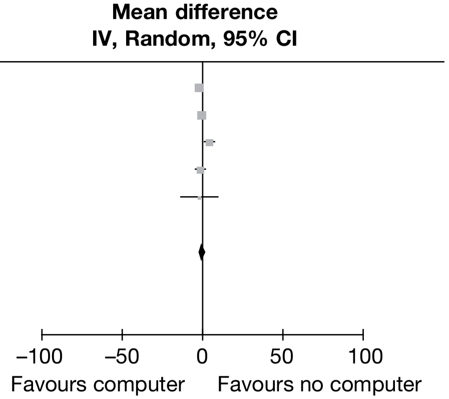


FIGURE 46 Meta-analysis MMPI – psychopathic deviation.

	Computer			No computer				
Study or subgroup	Mean	SD	Total	Mean	SD	Total	Weight (%)	Mean difference IV, Random, 95% CI
Biskin 1977[129]	14.108	8.714	37	15.4	8.593	45	24.1	–1.29 (–5.06 to 2.47)
White 1986[135]	15.86	6.443	50	14.74	7.82	50	43.3	1.12 (–1.69 to 3.93)
Honaker 1989[132]	56.95	11.796	40	56.55	9.476	40	15.5	0.40 (–4.29 to 5.09)
Lambert 1987[133]	22.19	11.3	38	23.45	9.54	37	15.3	–1.26 (–5.99 to 3.47)
Hart 1985[131]	70.2	17.3	10	80.1	14.1	10	1.8	–9.90 (–23.85 to 4.05)
Total (95% CI)			**175**			**182**	**100.0**	**–0.13 (–1.98 to 1.72)**

Heterogeneity: $\tau^2 = 0.00$; $\chi^2 = 3.28$, df = 4 ($p = 0.51$); $I^2 = 0\%$

Test for overall effect: $z = 0.14$ ($p = 0.89$)

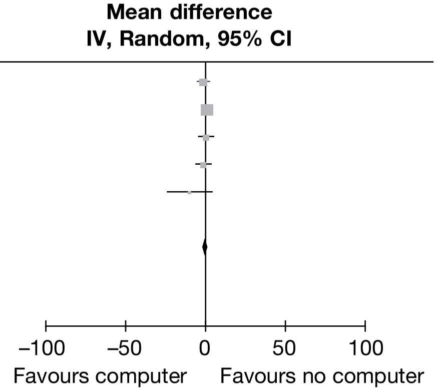


FIGURE 47 Meta-analysis MMPI – psychasthenia.

	Computer			No computer				
Study or subgroup	Mean	SD	Total	Mean	SD	Total	Weight (%)	Mean difference IV, Random, 95% CI
Biskin 1977[129]	15.432	9.72	37	16.178	9.263	45	21.6	–0.75 (–4.95 to 3.46)
White 1986[135]	13.86	6.491	50	13.98	7.389	50	51.4	–0.12 (–2.85 to 2.61)
Honaker 1989[132]	60.65	12.962	40	59.05	9.306	40	15.6	1.60 (–3.34 to 6.54)
Lambert 1987[133]	24.73	14.04	38	24.71	13.41	37	9.9	0.02 (–6.19 to 6.23)
Hart 1985[131]	75.6	19.5	10	86.1	16.7	10	1.5	–10.50 (–26.41 to 5.41)
Total (95% CI)			**175**			**182**	**100.0**	**–0.13 (–2.08 to 1.83)**

Heterogeneity: $\tau^2 = 0.00$; $\chi^2 = 2.19$, df = 4 ($p = 0.70$); $I^2 = 0\%$

Test for overall effect: $z = 0.13$ ($p = 0.90$)

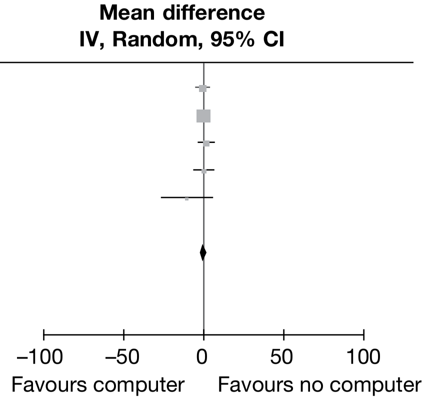


FIGURE 48 Meta-analysis MMPI – schizophrenia.

Chapter 5

Discussion

The theoretical review has resulted in a change in focus from modes as discrete entities to be compared with a focus on mode *features* as factors relating to the way in which responses on subjective outcomes are constructed by responders. These primary features come from the model previously suggested by Tourangeau *et al.*,[8] with additional potential features identified. These have then been tested in the results from a comprehensive systematic review of mode comparison studies in terms of their impact on *bias* and *precision*.

The results of the review of mode comparison studies clearly show that the impact of mode features is on bias rather than precision. Therefore, in planning a new study, choice of mode and features is unlikely to have a great impact on sample size considerations, but may have an impact between single-mode studies on interpretability of values of scores and within mixed-mode studies on the ability to simply combine scores collected under different mode features. This lack of an impact on precision also suggests that different mode features do not lead to differing degrees of end aversion bias or floor/ceiling effects.

The mode feature with the greatest impact, in terms of both magnitude (size of effect) and significance (strength of evidence), was mode of administration (interviewer or self). The choice of sensory stimuli (audio or visual or both) had a smaller impact (about half that of mode of administration), but this was just significant. Neither computerisation nor telephone primary features were significant in the main models, although there was some suggestion of a potential difference that had decreased over time when interaction terms were tested. This fits with previous suggestions that mode features relating to technologies initially lead to differences predominantly due to unfamiliarity in the responder with the technology and that as technologies move into common usage these differences reduce.

Of the additional or secondary features tested, differences in mode of delivery reached significance in some of the models, but with a smaller magnitude than either administration or sensory stimuli. This feature was proposed as tapping into the perceived legitimacy of data collection in terms of how an outcome measure was delivered or presented to a potential responder.

In addition to the theoretically derived mode features, a small number of potential mediators were included in the model. Very limited information was available from studies on these and, therefore, the only two considered were date of publication in relation to the introduction of new technologies and the number of items in a scale to relate to part of the construct of cognitive burden. This latter factor was significant, with single-item scales showing a greater degree of bias than multi-item scales.

Overall, the primary analysis of the mode comparison studies identified in the systematic review provides consistent evidence for the impact of two of the four theoretical mode features having an impact on the absolute mean difference (bias), but not on precision. However, the magnitude of these effects, when considered on a percentage scale is not great. Further exploration into the two most frequently occurring scales within our review (SF-36 and MMPI) showed mixed results. The analysis for these was carried out for each mode feature individually and, therefore, the potential findings for telephone and computer features may well be due to the confounding nature of also having a difference in administration. However, the estimation of the pooled limits of agreement for the within-person studies emphasises the potential impact of mode effects if the purpose of measurement is to consider an individual rather than a group. At a group level the main analysis indicates that on average the bias is significant, but relatively small in ES-terms. However, if the measure is to be used in clinical practice, for example, then the reliability of the assessment of the individual becomes important and the limits of agreement show how variable this can be.

The SF-36 collects data on health status and, therefore, represents an example of potentially more sensitive data (an antecedent feature). This may, therefore, increase the chances of satisficing (for example) and the importance of ensuring privacy (impersonality) in data collection. In the first analysis, addressing the use of a computer, there is a clear mode feature effect present for the mental health domains of the SF-36, but not the physical health domains. If (in these studies) computers served to enhance impersonality, these results are consistent with the framework. A challenge in interpreting such results is a general lack of detail relevant to the psychological appraisal processes available in published reports.

Strengths and weaknesses

This was a broad and comprehensive systematic review in terms of breadth of the published literature covered and the independence of discipline. Innovative approaches to designing search strategies have been tested and implemented in order to produce a search strategy with high levels of specificity. Grey literature was not looked into, given the large volume of evidence produced from published papers and abstracts. The search strategy only covered the period up until 2004; however, a considerable number of studies were identified and contributed to the analysis. Future updates could take a more focused approach on new and emerging technologies.

The focus on the mode features rather than the crude modes is consistent with a theoretical basis to the analysis and also takes further the exploration of the strengths of proposed relationships from theoretical models. The review of theory and discussion within a health framework provides researchers with an understanding of the potential impact of these features when designing their study.

We were not able to test all the potential mode features, with anonymity in responding being one where few data were provided in papers. There was also limited information on potential mediating factors such as cognitive burden and sensitivity questions. Overall, presentation of information was highly variable, and some approach to standardising reports of these types of study would be recommended in the future if they are to inform researchers on the portability of measures across mode features.

The presented framework directed the design of the data extraction sheet for the systematic review. This was most important in relation to the mode features and antecedent features. For example, variables (levels) included in the data extraction form were administration (self or interviewer), sensory channel (auditory, visual or both) and computer-assisted data collection (yes, no, don't know). Similarly, attempts were made to extract data related to the psychological appraisals. Thus, whether or not others were present at data collection and whether or not data collection ensured anonymity were both abstracted from studies. However, as expected, the availability of such data was limited in reviewed studies. Prospectively, a clear framework for conceptualising mode feature effects will be important for determining what data should be collected in empirical studies. Similarly, the analysis was guided by the framework, with key available variables from the framework included in the regression models. Hence, the initial regression model included all four mode features in the framework. Again, the lack of data about framework features recorded in published work limited, to some extent, the scale of this analysis for some cases.

Conclusions

Recommendations for researchers

Researchers need to be aware of the different mode features that could have an impact on their results when selecting a mode of data collection for subjective outcomes. If researchers use a mixture of modes within their study (commonly a change in mode if there is poor or non-response) then consideration needs to be given to ameliorating potential biases consequent to this and controlling for them in analysis.

The potential does exist for there to be simple correction factors developed; however, these are likely to be measure specific. In analysis of current mixed-mode studies, researchers cannot just assume that results are comparable where a difference in administration or sensory stimuli exists and need to either undertake sensitivity analyses or formally control for mode in the analysis.

Recommendations for future research (in priority order)

There is growing recognition within health research of the need to consider measurement equivalence across modes.[138] However, as evidenced in this review, there are already numerous studies considering a large number of outcome measures. However, these need to be reported in a standardised way to allow researchers to be able to make informed decisions about choice of mode with a particular outcome in a population. The development of reporting standards akin to PRISMA,[103] STROBE (Strengthening the Reporting of Observational Studies in Epidemiology)[139] or CONSORT (Consolidated Standards of Reporting Trials)[140] for mode comparison studies is urgently needed and could build on the quality assessment tool developed here.

Prospective empirical studies need to be more theoretically informed (i.e. designed to measure and test theoretically relevant components) and to report accordingly. Greater attempts within such research are needed to understand whether or not the mode features are actually mediated in the way hypothesised.

Further mode comparison studies are required, but these need to be experimentally designed to manipulate mode features and directly assess the impact. This is preferable to more studies comparing two modes at a relatively pragmatic level without consideration of those features. Studies need to give consideration to evaluation and direct testing of the impact of some of the mediators of mode effects, as the lack of data presented in papers in this review limited our ability to analyse this component.

Further primary studies need to be undertaken to evaluate the impact of mode features over time. There was a suggestion across studies that this occurred for 'new' technologies for data collection (telephone and computer), but the 'learning effect' for any mode over time will be important to evaluate further in order to inform studies with long-term follow-up over multiple time points. The potential biasing impact of this 'learning effect' over time could be seen in single-mode studies as well as mixed-mode ones.

The focus of this review has been on measurement for research purposes and, therefore, has focused predominantly on the impact of mode features on estimated effects at a group level. However, the increasing use of subjective patient-reported outcomes in clinical practice means that considerable further work is required to consider measurement equivalence and reliability of assessment of individuals rather than groups.

Chapter 6

Dissemination

Publication

1. Robling MR, Ingledew DK, Greene G, Sayers A, Shaw C, Sander L, *et al.* Applying an extended theoretical framework for data collection mode to health services research. *BMC Health Serv Res* 2010;**10**:180.

Oral presentations

1. Sayers A, on behalf of the MODE ARTS Team. A systematic literature review comparing multiple modes of survey administration: search strategy innovations. South West Society of Primary Care, Birmingham, UK, 2006.
2. Greene G, on behalf of the MODE ARTS Team. How does mode of survey administration affect the nature of the response provided? Some theoretical considerations. South West Society of Primary Care, Birmingham, UK, 2006.
3. Robling MR, on behalf of the MODE ARTS Team. Evaluating the impact of data collection mode upon response to subjective surveys: main results from the MODE ARTS systematic literature review. European Survey Research Association Biannual Conference, Prague, Czech Republic, 2007.
4. Robling MR, Hood K, Greene G, Sayers A, Ingledew DK, Russell IT, *et al.* Evaluating the impact of data collection mode upon response to subjective surveys: main results from the MODE ARTS systematic review. International Society for Quality of Life Research Annual Conference, Toronto, ON, Canada, 2007.

Poster presentations

1. Sayers A, on behalf of the MODE ARTS Team. A systematic literature review comparing multiple modes of survey administration: search strategy innovations. All Wales Systematic Review Symposium, Cardiff, UK, 2006.
2. Greene G, on behalf of the MODE ARTS Team. How does the modes of survey administration affect the response provided? All Wales Systematic Review Symposium, Cardiff, UK, 2006.

Projects/theses

1. Rhys Ivins. *Analysis of the Minnesota Multiphasic Personality Inventory*. Final Year Mathematics Undergraduate Project. Cardiff: Cardiff University; 2007
2. Adrian Sayers. *A comparison of different meta-analytic techniques: the use of triangulation in understanding the differences in response to surveys using different modes of administration*. MSc Project. London: London School of Hygiene & Tropical Medicine; 2007.

Acknowledgements

Lesley Sander contributed to the design of the project and a co-applicant on the bid.

Naomi Cadbury contributed to the data extraction and quality assessment.

Kristina Bennert, Jo Crocker, Marie Ann Durand, Monika Hare, Dyfed Hughes and Sew Tien Yeo extracted data from the foreign-language papers.

Jonathon Gillard supervised the industrial placement student and advised on analysis.

Amanda Iles and Aude Espinasse undertook data entry.

Contribution of authors

KH led on the design and management of the study, conducted the analysis of the mode comparisons studies and wrote the report.

MR led on the theory review and contributed to the design, management, paper review, interpretation of analysis and final report.

DI contributed to the theory review and to the design, management, interpretation of analysis and final report.

DG contributed to the analysis and final report.

GG undertook abstract and paper review, data extraction, quality assessment and contributed to the final report.

RI undertook data extraction, quality assessment, and contributed to the analysis and final report.

IR contributed to the design, management, interpretation of analysis and final report.

AS undertook the design and evaluation of the search strategy, abstract and paper review, and contributed to the final report.

CS contributed to the design, management, paper review, interpretation of analysis and final report.

JW contributed to the interpretation of analysis and final report.

All authors reviewed the final version of the manuscript.

References

1. US Food and Drug Administration (FDA). Guidance for Industry. Patient Reported Outcome Measures: use in medical product development to support labelling claims. 2009.
2. 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. *Heath Technol Assess* 2001;**5**(31).
3. Edwards PJ, Roberts I, Clarke M, DiGuiseppi C, Wentz R, Kwan I, *et al.* Methods to increase response to postal and electronic questionnaires. *Cochrane Database Syst Rev* 2009;**3**:MR000008.
4. Groves R, Couper M. *Nonresponse in household surveys*. New York, NY: Wiley; 1998.
5. Groves RM, Singer E, Corning A. Leverage–saliency theory of survey participation. *Publ Opin Q* 2000;**64**:299–308.
6. Tourangeau R. Cognitive aspects of survey measurement and mismeasurement. *Int J Public Opin Res* 2003;**15**:3–7.
7. Jobe JB. Cognitive psychology and self-reports: models and methods. *Qual Life Res* 2003;**12**:219–27.
8. Tourangeau R, Rips LJ, Rasinski K. *The psychology of survey response*. Cambridge: Cambridge University Press; 2000.
9. Tourangeau R, Smith TW. Asking sensitive questions: the impact of data collection mode, question format, and question context. *Publ Opin Q* 1996;**60**:275–304.
10. Honaker LM. The equivalency of computerised and conventional MMPI administration: a critical review. *Clin Psychol Rev* 1988;**8**:561–77.
11. Butler C, Kinnersley P, Hood K, Robling M, Prout H, Rollnick S, *et al.* The natural history of acute upper respiratory tract infection in children: a pragmatic randomised controlled trial cohort. *BMJ* 2003;**327**:1088–9.
12. Formica M, Kabbara K, Clark R, McAlindon T. Can clinical trials requiring frequent participant contact be conducted over the internet? Results from an online randomised controlled trial evaluating a topical ointment for herpes labialis. *J Med Internet Res* 2004;**6**:e6.
13. McAlindon T, Formica M, Kabbara K, LaValley M, Lehmer M. Conducting clinical trials over the internet: a feasibility study. *BMJ* 2003;**327**:484–7.
14. Steeh C, Kirgis N, Cannon B, DeWitt J. Are they really as bad as they seem? Nonresponse rates at the end of the twentieth century. *JOS* 2001;**17**:227–47.
15. Bloom DE. Technology, experimentation, and the quality of survey data. *Science* 1998;**280**:847–8.
16. Shih T-H, Fan X. Response rates and mode preferences in web-mail mixed-mode surveys: a meta-analysis. *Int J Internet Sci* 2007;**2**:59–82.
17. Shih T-H, Fan X. Comparing response rates in e-mail and paper surveys: a meta-analysis. *Educ Res Rev* 2009;**4**:26–40.
18. Groves RM, Fowler, Jr, FJ, Couper MP, Lepkowski JM, Singer E, Tourangeau R. *Survey methodology*. Hoboken, NJ: John Wiley & Sons; 2004.

19. Crow R, Gage H, Hampson S, Hart J, Kimber A, Storey L, *et al.* The measurement of satisfaction with healthcare: implications for practice from a systematic review of the literature. *Health Technol Assess* 2002;**6**(32).

20. Podsakoff PM, MacKenzie SB, Lee J-Y, Podsakoff NP. Common method biases in behavioural research: a critical review of the literature and recommended remedies. *J Appl Psychol* 2003;**88**:879–903.

21. Robling MR, Hood K, Greene G, Sayers A, Ingledew DK, Russell IT, *et al.* Evaluating the impact of data collection mode upon response to subjective surveys: main results from the MODEARTS systematic review. International Society for Quality of Life Research, 14th Annual Conference, Toronto, ON, Canada, 2007.

22. Birnbaum MH. Human research and data collection via the internet. *Ann Rev Psychol* 2004;**55**:803–32.

23. Reeve BB, Hays RD, Chang C-H, Perfetto EM. Applying item response theory to enhance health outcomes assessment. *Qual Life Res* 2007;**16**:1–3.

24. Walker J, Böhnke JR, Cerny T, Strasser F. Development of symptom assessments utilising item response theory and computer-adaptive testing: a practical method based on a systematic review. *Crit Rev Oncol Haematol* 2010;**73**:47–67.

25. Fayers PM. Applying item response theory and computer adaptive testing: the challenges for health outcomes assessment. *Qual Life Res* 2007;**16**:187–94.

26. Dillman DA, Christian LM. Survey mode as a source of instability in responses across surveys. *Field Methods* 2005;**17**:30–52.

27. Greenberg A, Manfield MN. On the reliability of mail questionnaires in product tests. *J Mark* 1957;**21**:342–5.

28. Sorensen S, Rylander R, Berglund K. Interviews and mailed questionnaires for the evaluation of annoyance reactions. *Environ Res* 1974;**8**:166–70.

29. Krosnick JA. Response strategies for coping with the cognitive demands of attitude measures in surveys. *Appl Cognitive Psychol* 1991;**5**:213–36.

30. Dubrovsky VJ, Kiesler S, Sethna BN. The Equalization Phenomenon: status effects in computer-mediated and face-to-face decision-making groups. *HCI* 1991;**6**:119–46.

31. Jones EF, Forrest JD. Underreporting of abortions in surveys of U.S. women: 1976 to 1988. *Demography* 1992;**29**:113–26.

32. Booth-Kewley S, Edwards JE, Rosenfeld P. Impression management, social desirability, and computer administration of attitude questionnaires: does the computer make a difference? *J Applied Psychol* 1992;**77**:562–6.

33. Turner CF, Ku L, Rogers SM, Lindberg LD, Pleck JH, Sonenstein FL. Adolescent sexual behaviour, drug use, and violence: increased reporting with computer survey technology. *Science* 1998;**280**:867–73.

34. Tourangeau R, Couper MP, Steiger DM. Humanising self-administered surveys: experiments on social presence in web and IVR surveys. *Comput Hum Behav* 2003;**19**:1–24.

35. Sproull L, Subramani M, Kiesler S, Walker JH, Waters K. When the interface is a face. *HCI* 1996;**11**:97–124.

36. Yates BT, Wagner JL, Suprenant LM. Recall of health-risky behaviours for the prior 2 or 4 weeks via computerised versus printed questionnaire. *Comput Hum Behav* 1997;**13**:83–110.

37. Smith TW. The impact of the presence of others on a respondent's answers to questions. *Int J Publ Opin Res* 1997;**9**:33–47.

38. Dillman DA. *Mail and internet surveys*. 2nd edn. New York, NY: John Wiley & Sons; 2000.

39. Krosnick JA. Survey research. *Annu Rev Psychol* 1999;**50**:537–67.

40. Krosnick JA, Holbrook AL, Berent MK, Carson RT, Hanemann WM, Kopp RJ, *et al.* The impact of 'no opinion' response options on data quality: non-attitude reduction or an invitation to satisfice? *Publ Opin Q* 2002;**66**:371–403.

41. Helgeson JG, Ursic ML. The decision process equivalency of electronic versus pencil-and-paper data collection methods. *Soc Sci Comp Rev* 1989;**7**:296–310.

42. Holbrook AL, Green MC, Krosnick JA. Telephone versus face-to-face interviewing of national probability samples with long questionnaires. *Publ Opin Q* 2003;**67**:79–125.

43. Kelly D, Harper DJ, Landau B. Questionnaire mode effects in interactive information retrieval experiments. *Inf Process Manag* 2008;**44**:122–41.

44. Krosnick JA, Alwin DF. An evaluation of cognitive theory of response-order effects in survey measurement. *Public Opin Q* 1987;**51**:201–19.

45. Jäckle A, Roberts C, Lynn P. *Telephone versus face-to-face interviewing: mode effects on data quality and likely causes. Report on phase II of the ESS-Gallup Mixed Mode Methodology Project*. Colchester: University of Essex; 2006.

46. Sudman S, Bradburn NM. *Response effects in surveys*. Chicago, IL: Aldine Publishing Company; 1974.

47. Sudman S, Bradburn NM. *Asking questions: A practical guide to questionnaire design*. San Francisco, CA: Jossey-Bass Publishers; 1982.

48. Paulhus DL. Two-component models of socially desirable responding. *J Pers Soc Psychol* 1984;**46**:598–609.

49. de Leeuw ED. *Data quality in mail, telephone and face to face surveys*. Amsterdam: TT-Publikaties; 1992.

50. Richman WL, Kiesler S, Weisband S, Drasgow F. A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires and interviews. *J Appl Psychol* 1999;**84**:754–75.

51. Dwight SA, Feigelson ME. A quantitative review of the effect of computerised testing on the measurement of social desirability. *Educ Psychol Meas* 2000;**60**:340–60.

52. Finegan JE, Allen NJ. Computerised and written questionnaires: are they equivalent? *Comput Hum Behav* 1994;**10**:483–96.

53. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF, *et al.* Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. *Lancet* 1999;**354**:1896–900.

54. Whitener EM, Klein HJ. Equivalence of computerised and traditional research methods: the roles of scanning, social environment, and social desirability. *Comput Hum Behav* 1995;**11**:65–75.

55. Shaeffer NC, Presser S. The science of asking questions. *Annu Rev Sociol* 2003;**29**:65–88.

56. Knowles ES, Condon CA. Why people say 'Yes': a dual-process theory of acquiescence. *J Pers Soc Psychol* 1999;**77**:379–86.

57. Jordan LA, Marcus AC, Reeder LG. Response styles in telephone and household interviewing: a field experiment. *Publ Opin Q* 1980;**44**:210–22.

58. Day H, Jankey S. Lessons from the literature: towards a holistic model of quality of life. In Renwick R, Brown I, Nagler M, editors. *Quality of life in health promotion and rehabilitation*. Thousand Oaks, CA: Sage Publications; 1996.

59. Schwartz CE, Rapkin BD. Reconsidering the psychometrics of quality of life assessment in light of response shift and appraisal. *Health Qual Life Outcomes* 2004;**2**:16.

60. Cote JA, Buckley R. Estimating trait, method, and error variance: generalising across 70 construct validation studies. *J Market Res* 1987;**24**:315–18.

61. Ghanem KG, Hutton HE, Zenilman JM, Zimba R, Erbelding EJ. Audio computer assisted self interview and face to face interview modes in assessing response bias among STD clinic patients. *Sex Transm Infect* 2004;**81**:421–5.

62. Hellard ME, Sinclair MI, Forbes AB, Fairley CK. Methods used to maintain a high level of involvement in a clinical trial. *J Epidemiol Community Health* 2001;**55**:348–51.

63. Chang B-H, Hendricks AM, Slawsky MT, Locastro JS. Patient recruitment to a randomised clinical trial of behavioural therapy for chronic heart failure. *BMC Med Res Methodol* 2003;**4**:8.

64. Hussain-Gambles M. South Asian patients' views and experiences of clinical trial participation. *Fam Pract* 2004;**21**:636–42.

65. Hewitt M. Attitudes towards interview mode and comparability of reporting sexual behaviour by personal interview and audio computer-assisted self-interviewing: analyses of the 1995 National Survey of Family Growth. *Sociol Methods Res* 2002;**31**:3.

66. Fitzpatrick R, Davey C, Buxton MJ, Jones DR. Evaluating patient-based outcome measures for use in clinical trials. *Health Technol Assess* 1998;**2**(14).

67. Dale O, Hagen KB. Despite technical problems personal digital assistants outperform pen and paper when collecting patient diary data. *J Clin Epidemiol* 2007;**60**:8–17.

68. Allenby A, Matthews J, Beresford J, McLachlan S. The application of computer touch-screen technology in screening for psychosocial distress in an ambulatory oncology setting. *Eur J Cancer Care* 2002;**11**:245–53.

69. Bushnell DM, Reilly MC, Galani C, Martin ML, Ricci J-F, Patrick DL, *et al.* Validation of electronic data capture of the Irritable Bowel Syndrome-Quality of Life measure, and Activity Impairment Questionnaire for Irritable Bowel Syndrome and the EuroQol. *Value Health* 2006;**9**:98–105.

70. Addington-Hall J. Research sensitivities to palliative care patients. *Eur J Cancer Care* 2002;**11**:220–4.

71. Marshall S, Haywood K, Fitzpatrick R. Impact of patient-reported outcome measures on routine practice: a structured review. *J Eval Clin Pract* 2006;**12**:559–68.

72. Dawson J, Doll H, Fitzpatrick R, Jenkinson C, Carr AJ. Routine use of patient reported outcome measures in healthcare settings. *BMJ* 2010;**340**:464–7.

73. Timmins N. Assessing patient care – NHS goes to the PROMS. *BMJ* 2008;**336**:1464–5.

74. Greenhalgh J, Meadows K. The effectiveness of the use of patient-based measures of health in routine practice in improving the process and outcomes of patient care: a literature review. *J Eval Clin Pract* 1999;**5**:410–16.

75. Espallargues M, Valderas JM. Provision of feedback on perceived health status to health care professionals: a systematic review of its impact. *Med Care* 2000;**38**:175–86.

76. Guyatt GH, Estwing Ferrans C, Halyard MY, Revicki DA, Symonds TL, Varricchio CG, *et al.* Exploration of the value of health-related quality-of-life information from clinical research and into clinical practice. *Mayo Clinic Proceedings* 2007;**82**:1229–39.

77. Gutteling JJ, Busschbach JJ, de Man RA, Darlington A-SE. Logistic feasibility of health related quality of life measurement in clinical practice: results of a prospective study in a large population of chronic liver patients. *Health Qual Life Outcomes* 2008;**6**:97.

78. Solari A. Role of health-related quality of life measures in the routine care of people with multiple sclerosis. *Health Qual Life Outcomes* 2005;**3**:16.

79. Greenhalgh J, Long AF, Flynn R. The use of patient reported outcome measures in routine clinical practice: lack of impact or lack of theory? *Soc Sci Med* 2005;**60**:833–43.

80. Department of Health. *Guidance on the routine collection of Patient Reported Outcome Measures (PROMs)*. London: Department of Health; 2008.

81. Varni JW, Burwinkle TM, Lane MM. Health-related quality of life measurement in paediatric clinical practice: an appraisal and precept for future research and application. *Health Qual Life Outcomes* 2005;**3**:34.

82. Engelen V, Haverman L, Koopman H, Schouten-van Meeteren N, Meijer-van den Bergh E, Vrijmoet-Wiersma J, *et al.* Development and implementation of a patient reported outcome intervention (QLIC-ON PROfile) in clinical paediatric oncology practice. *Patient Educ Couns* 2010:doi:10.1016/j.pec.2010.02.003.

83. Roberts C. *Mixing modes of data collection in surveys: a methodological review*. Southampton: ESRC National Centre for Research Methods; 2007.

84. Blumberg SJ, Luke JV, Cynamon ML. Telephone coverage and health survey estimates: evaluating the need for concern about wireless substitution. *Am J Public Health* 2006;**96**:926–31.

85. Chittleborough CR, Taylor AW, Baum FE, Hiller JE. Non-response to a life course socioeconomic position indicator in surveillance: comparison of telephone and face-to-face modes. *BMC Med Res Methodol* 2008;**8**:54.

86. Warner SL. Randomised response: a survey technique for eliminating evasive answer bias. *J Am Stat Assoc* 1965;**60**:63–9.

87. Lensvelt-Mulders GJLM, Hox JJ, van der Heijden PGM, Maas CJM. Meta-analysis of randomised response research: thirty-five years of validation. *Sociol Methods Res* 2005;**33**:319–48.

88. Rasinski KA, Visser PS, Zagatasky M, Rickett EM. Using implicit goal priming to improve the quality of self-report data. *J Exp Soc Psychol* 2005;**41**:321–7.

89. Groves RM, Fowler Jr FJ, Couper MP, Lepkowski JM, Singer E, Tourangeau R. *Survey methodology*. Wiley Series in Survey Methodology. London: John Wiley & Sons; 2004.

90. Carr E, Worth A. The use of telephone interview for research. *Nurs Times Res* 2001;**6**:511–24.

91. Hopewell S, McDonald S, Clarke M, Egger M. Grey literature in meta-analysis of randomised trials in health care interventions. *Cochrane Library* 2003;**1**.

92. Khan KS, Ter Riet G, Glanville J, Sowden AJ, Kleijnen J (editors). *Undertaking systematic reviews of research on effectiveness. CRD's guidance for carrying out or commissioning reviews.*

2nd edn. CRD Report No. 4. York: NHS Centre for Reviews and Dissemination (CRD), University of York; 2001.

93. de Leeuw ED, Hox JJ, Snijkers G. The effect of computer-assisted interviewing on data quality. A review. *J Market Res Soc* 1995;**37**:325.

94. Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *J Epidemiol Community Health* 1998;**52**:377–84.

95. Kmet L, Lee R, Cook L. *Standard quality assessment criteria for evaluating primary research papers from a variety of fields*. Edmonton, AB: Alberta Heritage Foundation for Medical Research; 2004.

96. Group TSoRT. A proposal for structured reporting of randomised controlled trials. The Standards of Reporting Trials Group. *JAMA* 1994;**272**:1926–31.

97. Anello C, Fleiss JL. Exploratory or analytic meta-analysis: should we distinguish between them? *J Clin Epidemiol* 1995;**48**:109–16.

98. van den Noorgate W, Onghena P. Multi-level meta-analysis: a comparison with traditional meta-analytic procedures. *Educ Psychol Meas* 2003;**63**:765–90.

99. Dunn G, Roberts C. Modelling method comparison data. *Stat Methods Med Res* 1999;**8**:161–78.

100. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res* 1999;**8**:135–60.

101. Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. *Statistician* 1983;**32**:307–17.

102. Williamson PR, Lancaster GA, Craig JV, Smuth RL. Meta-analysis of method comparison studies. *Stats Med* 2002;**21**:2013–25.

103. Moher D, Liberati A, Tetzlaff J, Altman DG, Group TP. Preferred Reporting Items for Systematic Reviews and Meta-analyses: The PRISMA statement. *BMJ* 2009;**339**:b2535. doi: 10.1136/bmj.b2535.

104. Bowling A, Bond M, Jenkinson C, Lamping DL. Short Form 36 (SF-36) Health Survey questionnaire: which normative data should be used? Comparisons between the norms provided by the Omnibus Survey in Britain, the Health Survey for England and the Oxford Healthy Life Survey. *J Public Health Med* 1999;**21**:255–70.

105. Berthelsen CL, Stilley KR. Automated personal health inventory for dentistry: a pilot study. *JADA* 2000;**131**:59–66.

106. Evans M, Kessler D, Lewis G, Peters TJ, Sharp D. Assessing mental health in primary care research using standardised scales: can it be carried out over the telephone? *Psychol Med* 2004;**34**:157–62.

107. Allison T, Ahmad T, Brammah T, Symmons D, Urwin M. Can findings from postal questionnaires be combined with interview results to improve the response rate among ethnic minority populations? *Ethn Health* 2003;**8**:63–9.

108. Ooijen MV, Ivens UI, Johansen C, Skov T. Comparison of a self-administered questionnaire and a telephone interview of 146 Danish waste collectors. *Am J Ind Med* 1997;**31**:653–8.

109. Cohen J. *Statistical power analysis for the behavioural sciences.* 2nd edn. Hillsdale, NJ: Lawrence Erlbaum; 1988.

110. Ware J, Kosinski M, Dewey J. *How to score version two of the SF-36 Health Survey*. Lincoln, RI: QualityMetric Incorporated; 2000.

111. Amodei N, Katerndahl DA, Larme AC, Palmer R. Interview versus self-answer methods of assessing health and emotional functioning in primary care patients. *Psychol Rep* 2003;**92**:937–48.

112. Jones D, Kazis L, Lee A, Rogers W, Skinner K, Cassar L, *et al.* Health status assessments using the Veterans SF-12 and SF-36: methods for evaluating outcomes in the Veterans Health Administration. *J Ambul Care Manage* 2001;**24**:68–86.

113. Lyons RA, Wareham K, Lucas M, Price D, Williams J, Hutchings HA. SF-36 scores vary by method of administration: implications for study design. *J Public Health Med* 1999;**21**:41–5.

114. Perkins JJ, Sanson-Fisher RW. An examination of self- and telephone-administered modes of administration for the Australian SF-36. *J Clin Epidemiol* 1998;**51**:969–73.

115. Saleh KJ, Radosevich DM, Kassim RA, Moussa M, Dykes D, Bottolfson H, *et al.* Comparison of commonly used orthopaedic outcome measures using palm-top computers and paper surveys. *J Orthop Res* 2002;**20**:1146–51.

116. Unruh M, Yan G, Radeva M, Hays RD, Benz R, Athienites NV, *et al.* Bias in assessment of health-related quality of life in a hemodialysis population: a comparison of self-administered and interviewer-administered surveys in the HEMO study. *J Am Soc Nephrol* 2003;**14**:2132–41.

117. van Campen C, Sixma H, Kerssens JJ, Peters L. Comparisons of the costs and quality of patient data collection by mail versus telephone versus in-person interviews. *Eur J Public Health* 1998;**8**:66–70.

118. Weinberger M, Oddone EZ, Samsa GP, Landsman PB. Are health-related quality-of-life measures affected by the mode of administration? *J Clin Epidemiol* 1996;**49**:135–40.

119. Abdoh A, Krousel-Wood MA, Re RN. Validity And Reliability Assessment Of An Automated Telephone Survey System (208). *Congress Epidemiol Abstr* 2001:S87.

120. Bliven BD, Kaufman SE, Spertus JA. Electronic collection of health-related quality of life data: validity, time benefits, and patient preference. *Qual Life Res* 2001;**10**:15–22.

121. Caro JJ, Caro I, Caro J, Wouters F, Juniper EF. Does electronic implementation of questionnaires used in asthma alter responses compared to paper implementation? *Qual Life Res* 2001;**10**:683–91.

122. Molitor F, Kravitz RL, To Y, Fink A. Methods in survey research: evidence for the reliability of group administration vs personal interviews. *Am J Public Health* 2001;**91**:826–7.

123. Revicki DA, Tohen M, Gyulai L, Thompson C, Pike S, Davis-Vogel A, *et al.* Telephone versus in-person clinical and health status assessment interviews in patients with bipolar disorder. *Harvard Rev Psychiatr* 1997;**5**:75–81.

124. Ryan JM, Corry JR, Attewell R, Smithson MJ. A comparison of an electronic version of the SF-36 General Health Questionnaire to the standard paper version. *Qual Life Res* 2002;**11**:19–26.

125. Weinberger M, Nagle B, Hanlon JT, Samsa GP, Schmader K, Landsman PB, *et al.* Assessing health-related quality of life in elderly outpatients: telephone versus face-to-face administration. *J Am Geriatr Soc* 1994;**42**:1295–9.

126. Wilson AS, Kitas GD, Carruthers DM, Reay C, Skan J, Harris S, *et al.* Computerised information-gathering in specialist rheumatology clinics: an initial evaluation of an electronic version of the Short Form 36. *Rheumatology* 2002;**41**:268–73.

127. Kazis L. The Veterans SF-36 health status questionnaire: development and application in the Veteran's Health Administration. *Med Outcomes Trust Monit* 2000;**5**(1).

128. Graham JR. *MMPI-2: Assessing personality and psychopathology.* 4th edn. New York, NY: Oxford University Press; 2006.

129. Biskin BH, Kolotkin RL. Effects of computerised administration on scores on the Minnesota Multiphasic Personality Inventory. *Appl Psychol Meas* 1977;**1**:543–9.

130. Evan WM, Miller JR. Differential effects on response bias of computer vs. conventional administration of a social science questionnaire: An exploratory methodological experiment. *Behav Sci* 1969;**14**:216–27.

131. Hart RR, Goldstein MA. Computer-assisted psychological assessment. *Comput Hum Serv* 1985;**1**:69–75.

132. Honaker L, Harrell TH, Buffaloe JD. Equivalency of Microtest computer MMPI administration for standard and special scales. *Comput Hum Behav* 1988;**4**:323–37.

133. Lambert ME, Andrews RH, Rylee K, Skinner JR. Equivalence of computerised and traditional MMPI administration with substance abusers. *Comput Hum Behav* 1987;**3**:139–43.

134. Locke SD, Gilbert BO. Method of psychological assessment, self-disclosure, and experiential differences: A study of computer, questionnaire, and interview assessment formats. *J Soc Behav Pers* 1995;**10**:255–63.

135. White DM, Clements CB, Fowler RD. A comparison of computer administration with standard administration of the MMPI. *Comput Hum Behav* 1985;**1**:153–62.

136. Pinsoneault TB. Equivalency of computer-assisted and paper-and-pencil administered versions of the Minnesota Multiphasic Personality Inventory-2. *Comput Hum Behav* 1996;**12**:291–300.

137. Schuldberg D. The MMPI is less sensitive to the automated testing format than it is to repeated testing: Item and scale effects. *Comput Hum Behav* 1988;**4**:285–98.

138. Coons SJ, Gwaltney CJ, Hayes RD, Lundy JJ, Sloan JA, Revicki DA, *et al.* Recommendations on evidence needed to support measurement equivalence between electronic and paper-based patient reported outcome (PRO) measures: ISPOR ePRO good research practices task force report. *Value Health* 2009;**12**:419–29.

139. Elm EV, Altman DG, Eggar M, Pocock SJ, Gotzsche PC, Vandenbroucke JP, *et al.* The Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting of observational studies. *J Clin Epidemiol* 2008;**61**:344–9.

140. Shulz KF, Altman DG, Moher D, Group C. *CONSORT* 2010 Statement: updated guidelines for reporting parallel group randomised trials. *Ann Intern Med* 2010;152.

141. Albaum GS, Evangelista F, Medina N. Role of response behaviour theory in survey research: a cross-national study. *J Business Res* 1998;**42**:115–25.

142. Willis G, Royston P, Bercini D. The use of verbal report methods in the development and testing of survey questionnaires. *Appl Cogn Psychol* 1991;**5**:251–67.

143. Collins D. Pretesting survey instruments: an overview of cognitive methods. *Qual Life Res* 2003;**12**:229–38.

144. McColl E, Meadows K, Barofsky I. Cognitive aspects of survey methodology and quality of life assessment. *Qual Life Res* 2003;**12**:217–18.

145. Tourangeau R. Survey research and societal change. *Ann Rev Psychol* 2004;**55**:775–801.

146. 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. *Soc Sci Med* 1991;**33**:1303–8.

147. Weir P, Beri S. A comparison and evaluation of two survey data collection methodologies: CATI vs mail. *Proc Surv Res Methods Sec Am Stat Assoc* 2000:388–93.

148. Edwards P, Roberts I, Clarke M, DiGuiseppi C, Pratap S, Wentz R, *et al.* Increasing response rates to postal questionnaires: systematic review. *BMJ* 2002;**324**:1183–7.

149. Krysan M, Schuman H, Scott LJ, Beatty P. Response rates and response content in mail versus face-to-face surveys. *Publ Opin Q* 1994;**58**:381–99.

150. Damiano A, McGrath M, William M, Snyder C, LeWitt P, Reyes P, *et al.* Evaluation of a measurement strategy for Parkinson's Disease: assessing patient health-related quality of life. *Qual Life Res* 2000;**9**:87–100.

Appendix 1

Search strategy

TABLE 21 Databases searched

Databases	Indexed: from-2004	Database provider
Health		
AMED	1985	Ovid
BNI	1985	Ovid
CINAHL	1982	Ovid
EMBASE	1980	Ovid
MEDLINE	1966	Ovid
Old MEDLINE	1950–1965	Ovid
Evidence-based medicine		
ACP Journal Club		Ovid
CCRCT		Ovid
CDSR		Ovid
DARE		Ovid
Social sciences		
ASSIA	1987	CSA
PsycINFO	1806	Ovid
SCI	1970	WoK
Social service abstracts	1979	CSA
Sociological abstracts	1952	CSA
SSCI	1970	WoK
Economics		
EconLit	1969	Ovid
Other		
SPORTDiscus	1830	Ovid
Hand-searching		
ASA – Survey Research Methods	1978	ASA

ACP, American College of Physicians; AMED, Allied and Complementary Medicine Database; ASA, American Statistical Association; ASSIA, Applied Social Sciences Index and Abstracts; BNI, British Nursing Index; CCRCT, Cochrane Central Register of Controlled Trials; CDSR, Cochrane Database of Systematic Reviews; CINAHL, Cumulative Index to Nursing and Allied Health Literature; CSA, CSA Illumina; DARE, Database of Abstracts of Reviews of Effects; EMBASE, Excerpta Medica Database; SCI, Science Citation Index; SSCI, Social Science Citation Index; WoK, Web of Knowledge; Ovid, Ovid Technologies.

Final search strategy

Search string 0–1

Finalised search strategy implemented in all databases allowing for changes in field codes and thesaurus terms.

1. Computer$.ti,ab OR Mini Computer$ OR Mini-Computer$ OR Minicomputer$ OR Micro Computer$ OR Micro-Computer$ OR Microcomputer$ OR Multi Media OR Multi-Media OR Multimedia OR ACAPI OR CAPI OR CASI OR CACPI OR Touch Screen$ OR Touch-Screen$ OR Touchscreen$ OR Portable Computer$ OR Portable-Computer$ OR Portablecomputer$ OR PDA OR PDAs OR PDA's OR Personal Digital Assistant$ OR Personal-Digital- Assistant$ OR Personaldigitalassistant$ OR Personal-Digital Assistant$ OR Personal Digital-Assistant$ OR Pocket PC$ OR Pocket-PC$ OR Pocketpc$ OR Palm OR Psion$ OR Pocket Computer$ OR Pocket-Computer$ OR Lap Top$ OR Lap-Top$ OR Laptop$ OR Notebook$ OR Note Book$ OR Note-Book$ OR Pen Tablet$ OR Pen-Tablet$ OR Pentablet$ OR Virtual OR Interactive OR E mail$ OR E-mail$ OR Email$ OR Electronic Mail$ OR Electronic-Email$ OR Electronicmail$ OR Electronic Diar$ OR Electronic-Diar$ OR Electronicdiar$ OR HHC OR CAI OR ACASI OR PTC OR Palm Top OR Palm-Top OR Palmtop OR E-Diary OR Ediary OR Automated OR [Technology Assisted Thesaurus Terms]
2. World-Wide-Web OR World-Wide Web OR World Wide Web OR Worldwide Web OR WWW OR On Line OR Online OR On-line OR Internet$ OR Inter-Net$ OR Inter Net$ OR Intranet$ OR Intra-Net$ OR Intra Net$ OR Web Based OR Web-Based OR Webbased
3. Offline OR Off Line OR Off-Line OR Unplugged OR Un Plugged OR Un-Plugged
4. Paper and Pen$ OR Pen$and Paper OR Pen Paper OR Pen-Paper OR Paper Pen OR Paper-Pen OR Paper Based OR Paper-Based OR Paperbased OR Papi OR Self Answer$ OR Self-Answer$ OR Selfanswer$ OR Self Administ$ OR Self-Administ$ OR Selfadminist$ OR Self Complete$ OR Self-Complete$ OR Selfcomplete$ OR Self Interview$ OR Self- Interview$ OR Selfinterview$ OR Self Report$ OR Self-Report$ OR Selfreport$ OR Diary OR Diaries OR Mail$ OR Posted OR Postal OR Questionnaire$ OR Paper/pencil OR Paper/Pencil OR PPQ OR P&P OR Snail Mail OR Snail-Mail OR Snailmail OR Journal OR Log OR SAQ OR Self Disclosure
5. Facsimile OR Fax OR Telefax OR Telefacsmile
6. Telephone$ OR Cellular Phone OR Cellular-Phone$ OR Cellularphone Phone$ OR CATI OR CACI
7. Face to Face OR Facetoface OR Face-to-Face OR Interview$ OR Door to Door OR Door-to-Door OR Door-to Door OR Door to-Door OR Curb Side OR Curb-Side OR Curbside OR Face-to Face OR Face to-Face OR Person to Person OR Person-to-Person OR Person-to Person OR Person to-Person OR FTFI OR FTF OR F2F
8. Mode OR Modes OR Modal
9. Video$
10. ACAPI OR ACASI OR Automated OR CACI OR CACPI OR CAI OR CAPI OR CASI OR CATI OR Cellular Chone$ OR Cellularphone$ OR Cellular-Phone$ OR Computer$ OR Curb Side OR Curb-Side OR Curbside OR Diary OR Diaries OR Door to Door OR Door-to-Door OR Door-to Door OR Door to-Door OR E mail$ OR E-mail$ OR Email$ OR Electronic$ OR E-Diary OR Ediary OR Face to Face OR Face-to-Face OR Facetoface OR Face-to Face OR Face to-Face OR Facsimile OR Fax OR Telefax OR Telefacsimile OR FTFI OR FTF OR F2F OR HHC OR Inter Net$ OR Inter-Net$ OR Internet$ OR Interactive OR Interview$ OR Intra Net$ OR Intra-Net$ OR Intranet$ OR Journal OR Lap Top$ OR Lap-Top$ OR Laptop$ OR Log OR Mail$ OR Medium OR Method$ OR Micro Computer$ OR Micro-Computer$ OR Microcomputer$ OR Mini Computer$ OR Mini-Computer$ OR Minicomputer$ OR Modal OR Mode OR Modes OR Multi Media OR Multi-Media OR Multimedia OR Note Book$ OR Note-Book$ OR Notebook$ OR Offline OR Off Line OR Off-Line OR On Line OR On-Line OR Online OR Palm OR Paper$ OR Pen OR Pencil$ OR Pens OR Paper/pencil OR Paper/Pencil OR PAPI OR PC OR PDA OR PDAs OR PDA's OR Pen tablet$ OR Pen-Tablet$ OR Pentablet$ OR Person to Person OR Person-to-Person OR Person-to Person OR Person to-Person OR Personal Digital Assistant$ OR Personal-Digital-Assistant$ OR Personaldigitalassistant$ OR Personal Digital-Assistant$ OR Personal-Digital Assistant$ OR Phone$ OR Pocket Computer$ OR Pocket-Computer$ OR Pocket PC$ OR

Pocket-PC$ OR Pocketpc$ OR Portable Computer$ OR Portable-Computer$ OR Portable Computer$ OR Postal OR Posted OR PPQ OR P&P OR Psion$ OR PTC OR Palm Top OR Palm-Top OR Palmtop OR Questionnaire$ OR SAQ OR Self Administ$ OR Self-Administ$ OR Selfadminist$ OR Self Answer$ OR Self-Answer$ OR Selfanswer$ OR Self Complet$ OR Self-Complet$ OR Selfcomplet$ OR Self Interview$ OR Self-Interview$ OR Selfinterview$ OR Self Disclosure OR Self Report$ OR Self-Report$ OR Selfreport$ OR Snail Mail OR Snail-Mail OR Snailmail OR Technology OR Telephone$ OR Touch Screen$ OR Touch-Screen$ OR Touchscreen$ OR Traditional OR Unplugged OR Un Plugged OR Un-Plugged OR Valid$ OR Video$ OR Virtual OR Web OR Webbased OR World-Wide-Web OR WWW

11. Alternat$ OR Blind$ OR Compar$ OR Concurrence OR Consist$ OR Contrast$ OR Control$ OR Cross Over OR Crossover OR Cross-Over OR Differ$ OR Error$ OR Evaluat$ OR Feasibility OR Group$ OR Mask$ OR Method.kw OR Methodolog$ OR Random$ OR Reliab$ OR Reproducibility of Results OR Sensitivity OR Specificity OR Survey OR Valid$ OR Versus$ OR Vs OR V's
12. Administration$ OR Assessment$ OR Data Collect$ OR Diaries OR Diary OR Examination$ OR Interview$ OR Questionnaire$ OR Screen$ OR Self-report$ OR Survey$ OR Test$ OR [Comparative Thesaurus Terms]
13. 1 AND (OR/2 – 9)
14. 2 AND (OR/3 – 9)
15. 3 AND (OR/4 – 9)
16. 4 AND (OR/5 – 9)
17. 5 AND (OR/6 – 9)
18. 6 AND (OR/7 – 9)
19. 7 AND (OR/8 – 9)
20. 8 AND 9
21. OR/13 – 20
22. 10 AND 11 AND 12 AND 21
23. Limit 22 to Human
24. Limit 23 to yr = [Start Date – 2004]

Appendix 2

Data extraction sheets

Datasheet 1: full-paper initial screen

PAPER ID:

Extracted by: Date of extraction:

Does this paper compare 2 or more modes of data collection*? Y / N (if N, then STOP)

Modes compared? ____

Levels of reporting:

Response rates Y / N Data quality Y / N

Is the measurement of the same construct compared across different modes? Y / N (if N, then STOP)

Does the comparison involve a diagnostic interview? Y / N

Measure	Construct	Subjective*, self-report? (? = for discussion, judgment = N*)
		Y / N / ?
		Y / N / ?
		Y / N / ?
		Y / N / ?
		Y / N / ?
		Y / N / ?
		Y / N / ?
		Y / N / ?
		Y / N / ?

WITHIN DESIGN

- is the mode effect confounded by time of data collection* Y / N

BETWEEN DESIGN

- is the mode effect confounded by the sampling strategy* Y / N

Notes:

DECISION: IN OUT FOR DISCUSSION

* SEE DEFINITIONS FOR CLARIFICATION

DEFINITIONS

MODE COMPARISON

A mode comparison study is one in which the same construct is measured (either with or without the same tool administered) in two different modes, the scores are computed in the same way, and the scores are (or can be) compared.

SUBJECTIVE

A subjective construct is one that is only accessible through an individual's subjective self-report (whether the self-report is recorded by the individual or by an interviewer or other person).

JUDGMENT

A study involves a judgment if the performance on the measure informs a judgment defined by an external source e.g. a diagnosis, rather than the actual score derived from the measure.

WITHIN DESIGN

Confounds with the time of data collection relate to studies in which:-

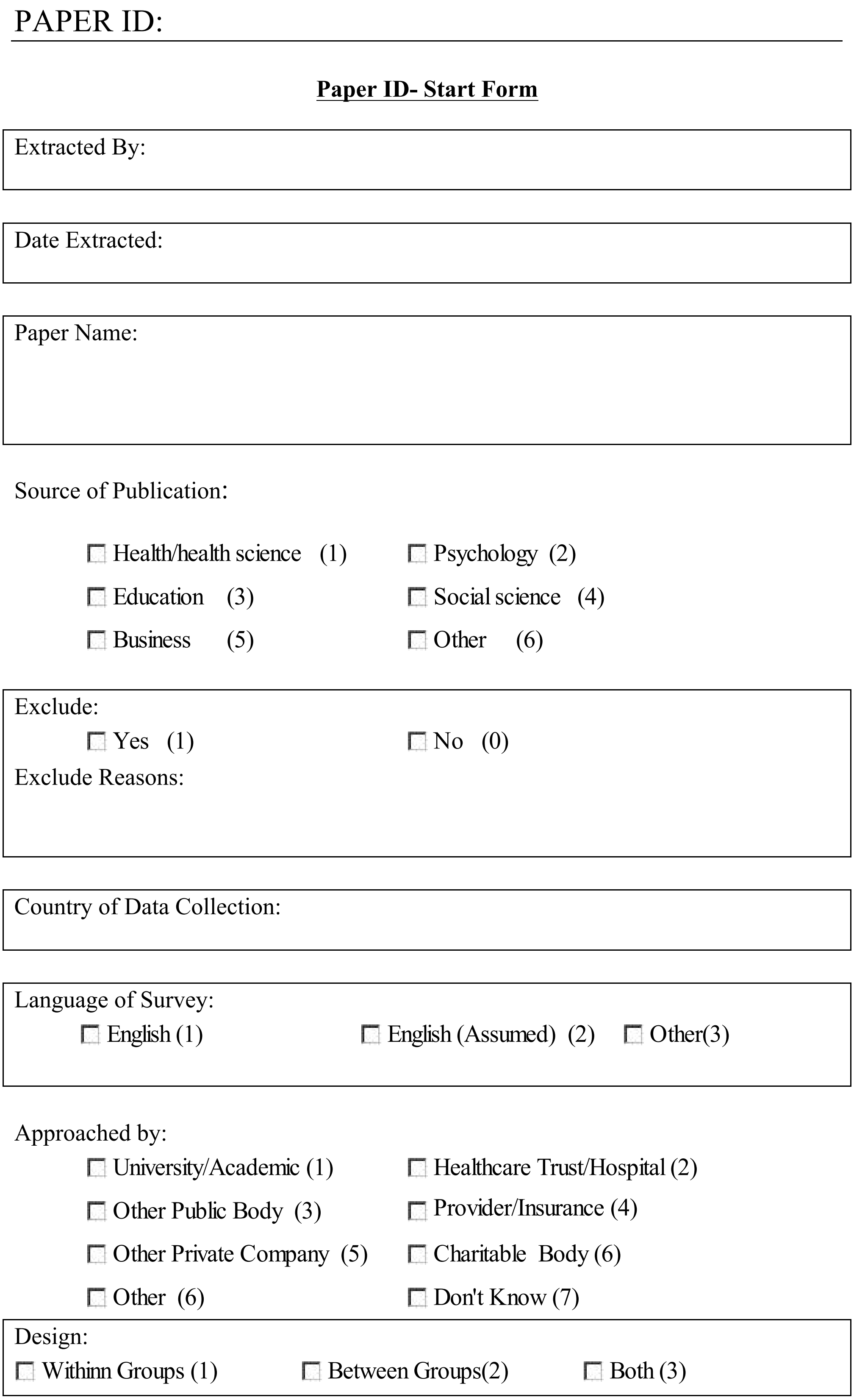
a) The use of two different collection methods that are not collecting data relating to the same time e.g. the use of a daily diary vs. a bi-weekly telephone interview

BETWEEN DESIGN

Confounds in the sampling strategy are

a) when the sampling frame for groups are determined by different methods, e.g. door-to-door interviews within a small community (city block) vs. random digit dialling of a much larger community (city)

Datasheet 2: paper ID – paper information



PAPER ID:

Paper ID- Start Form

Extracted By:

Date Extracted:

Paper Name:

Source of Publication:

□ Health/health science (1) □ Psychology (2)

□ Education (3) □ Social science (4)

□ Business (5) □ Other (6)

Exclude:

□ Yes (1) □ No (0)

Exclude Reasons:

Country of Data Collection:

Language of Survey:

□ English (1) □ English (Assumed) (2) □ Other(3)

Approached by:

□ University/Academic (1) □ Healthcare Trust/Hospital (2)

□ Other Public Body (3) □ Provider/Insurance (4)

□ Other Private Company (5) □ Charitable Body (6)

□ Other (6) □ Don't Know (7)

Design:

□ Withinn Groups (1) □ Between Groups(2) □ Both (3)

Datasheet 3: sample and demographics data

PAPER ID:

Sample and Demographics Data

Population:

Site of Data Collection:

Data Collection Team:

Date of Data Collection:

Time Frame:
Units:
□ Hours (1) □ Days (2) □ Months (3) □ Years (4)

Sampling Strategy:
□ RDD (1) □ Targeted (2) □ Targeted - Clinic Lists (3) □ Convenience (4)
□ Random (5) □ Random Stratified (6) □ Systematic (7) □ Stratified (8)

Target Time Gap (T1-T2):
□ Hours (1) □ Days (2) □ Months (3) □ Years (4)
Justification of Time Gap:

Mean Achieved Time Gap:

SD Achieved Time Gap:

Range of Achieved Time:

Order Allocation:
□ All the Same (1) □ Random (2) □ Systematic (3) □ Sampling (4) □ Other (5)

Group Allocation Other:

Population Description:

Personality Description:

PAPER ID:

No. of Modes Compared:

N	N=Females	Age	SD Age

Ethnicity:

Educational Status:

SES:

Employment:

Notes:

Rewards:

□ Yes (1) □ No (0)

Details:

Relevence:

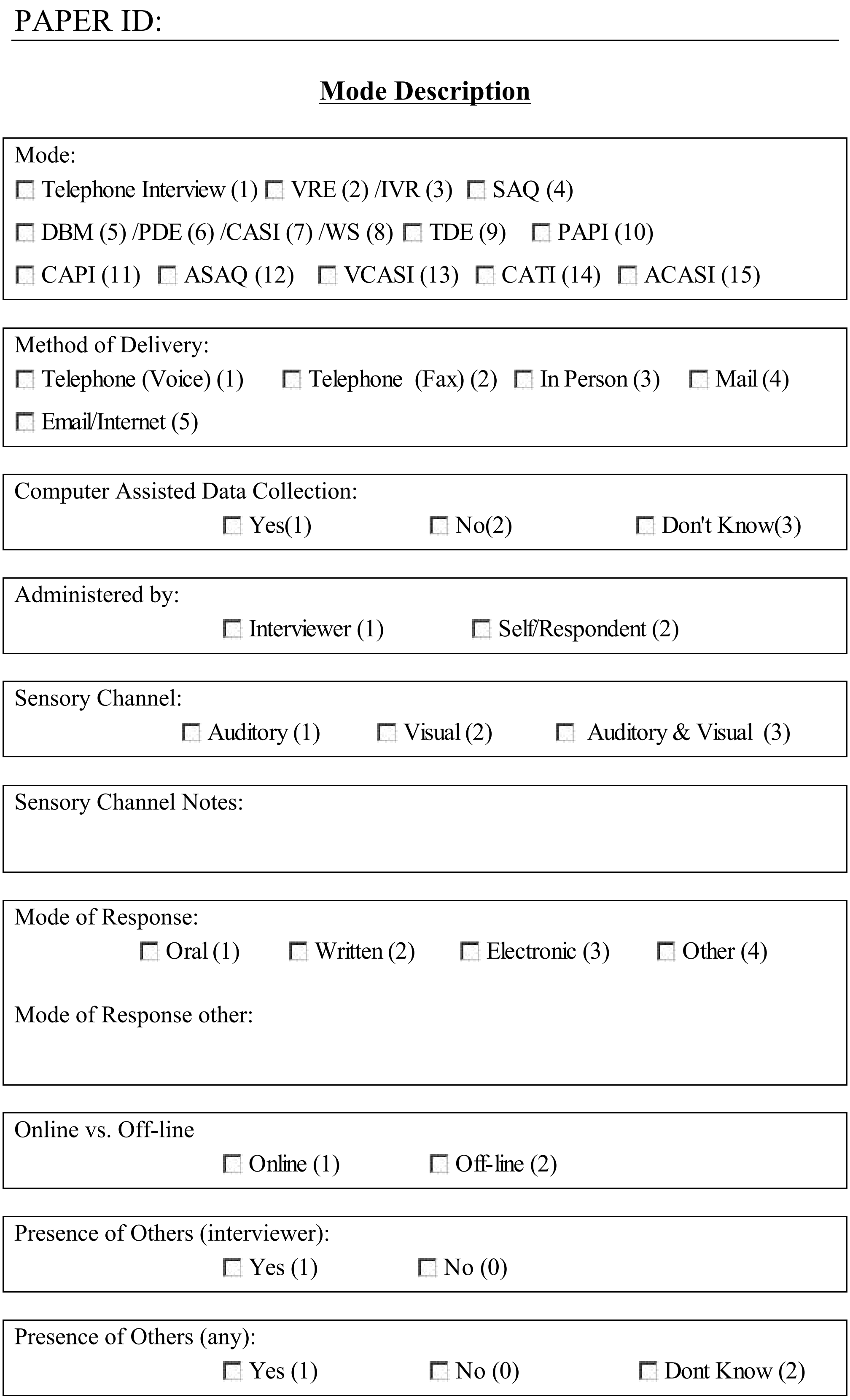
□ Yes (1) □ No (0) □ Not Sure (Details) (2)

Details:

Knowledge of Repeated Design:

□ Yes (1) □ No (0) □ Not Sure (2)

Datasheet 4: mode description



PAPER ID:

Mode Description

Mode:
□ Telephone Interview (1) □ VRE (2) /IVR (3) □ SAQ (4)
□ DBM (5) /PDE (6) /CASI (7) /WS (8) □ TDE (9) □ PAPI (10)
□ CAPI (11) □ ASAQ (12) □ VCASI (13) □ CATI (14) □ ACASI (15)

Method of Delivery:
□ Telephone (Voice) (1) □ Telephone (Fax) (2) □ In Person (3) □ Mail (4)
□ Email/Internet (5)

Computer Assisted Data Collection:
□ Yes(1) □ No(2) □ Don't Know(3)

Administered by:
□ Interviewer (1) □ Self/Respondent (2)

Sensory Channel:
□ Auditory (1) □ Visual (2) □ Auditory & Visual (3)

Sensory Channel Notes:

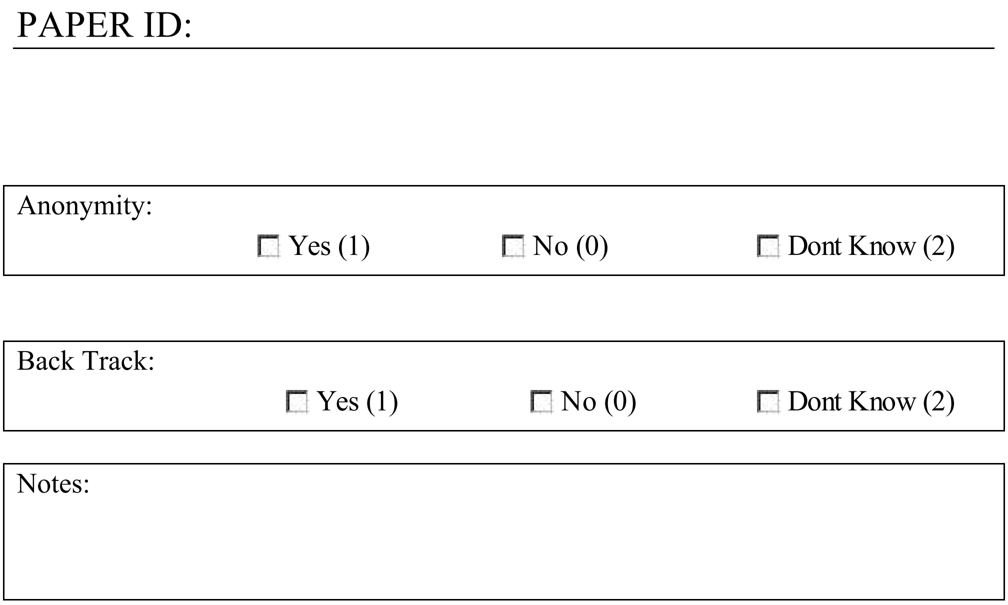
Mode of Response:
□ Oral (1) □ Written (2) □ Electronic (3) □ Other (4)

Mode of Response other:

Online vs. Off-line
□ Online (1) □ Off-line (2)

Presence of Others (interviewer):
□ Yes (1) □ No (0)

Presence of Others (any):
□ Yes (1) □ No (0) □ Dont Know (2)



PAPER ID:

Anonymity:

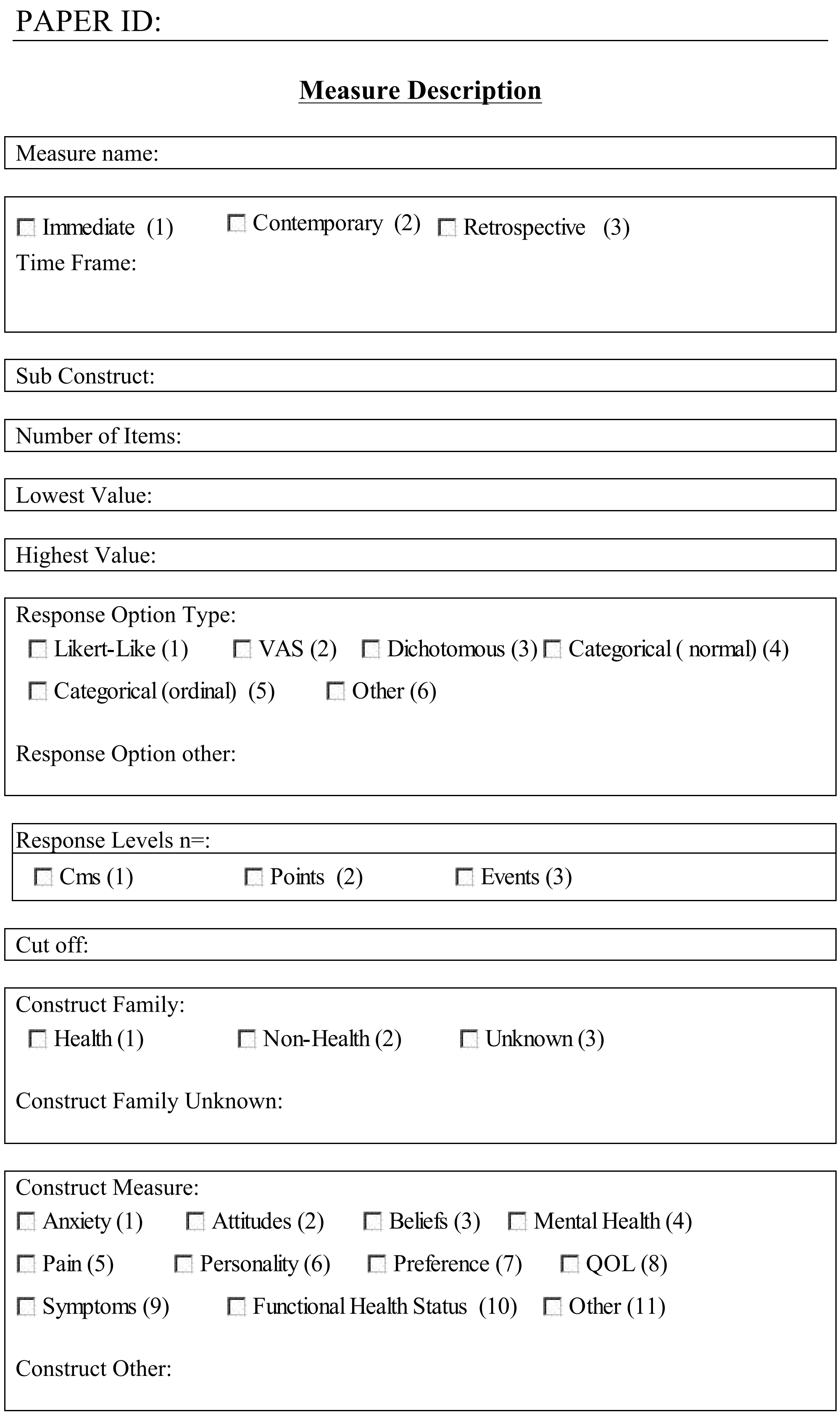
□ Yes (1) □ No (0) □ Dont Know (2)

Back Track:

□ Yes (1) □ No (0) □ Dont Know (2)

Notes:

Datasheet 5: measure description



PAPER ID:

Measure Description

Measure name:

□ Immediate (1) □ Contemporary (2) □ Retrospective (3)

Time Frame:

Sub Construct:

Number of Items:

Lowest Value:

Highest Value:

Response Option Type:

□ Likert-Like (1) □ VAS (2) □ Dichotomous (3) □ Categorical (normal) (4)

□ Categorical (ordinal) (5) □ Other (6)

Response Option other:

Response Levels n=:

□ Cms (1) □ Points (2) □ Events (3)

Cut off:

Construct Family:

□ Health (1) □ Non-Health (2) □ Unknown (3)

Construct Family Unknown:

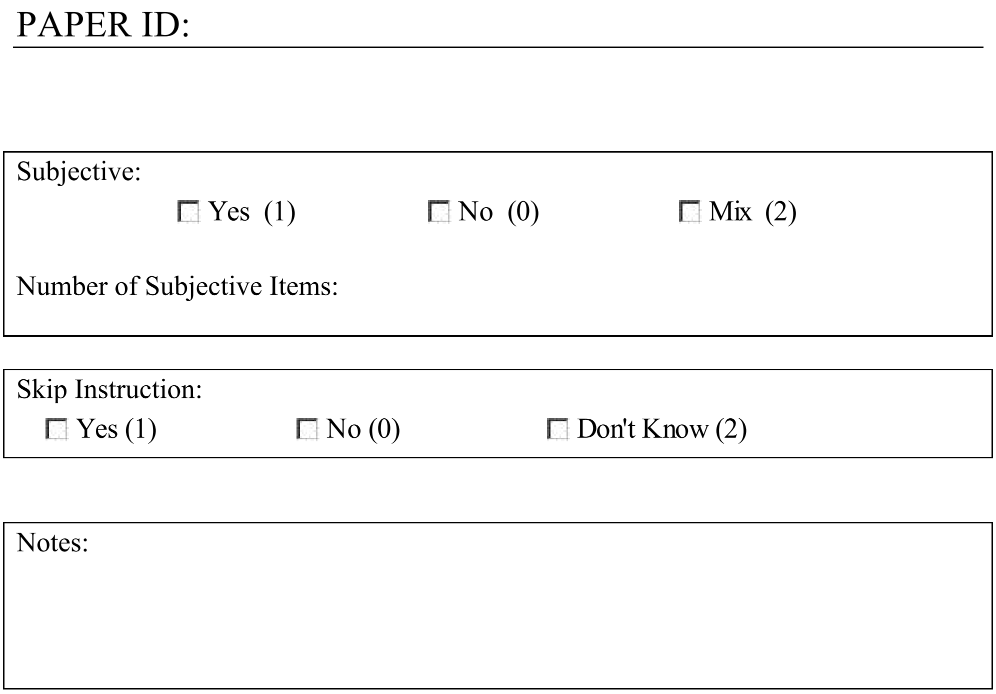
Construct Measure:

□ Anxiety (1) □ Attitudes (2) □ Beliefs (3) □ Mental Health (4)

□ Pain (5) □ Personality (6) □ Preference (7) □ QOL (8)

□ Symptoms (9) □ Functional Health Status (10) □ Other (11)

Construct Other:



PAPER ID:

Subjective:

□ Yes (1) □ No (0) □ Mix (2)

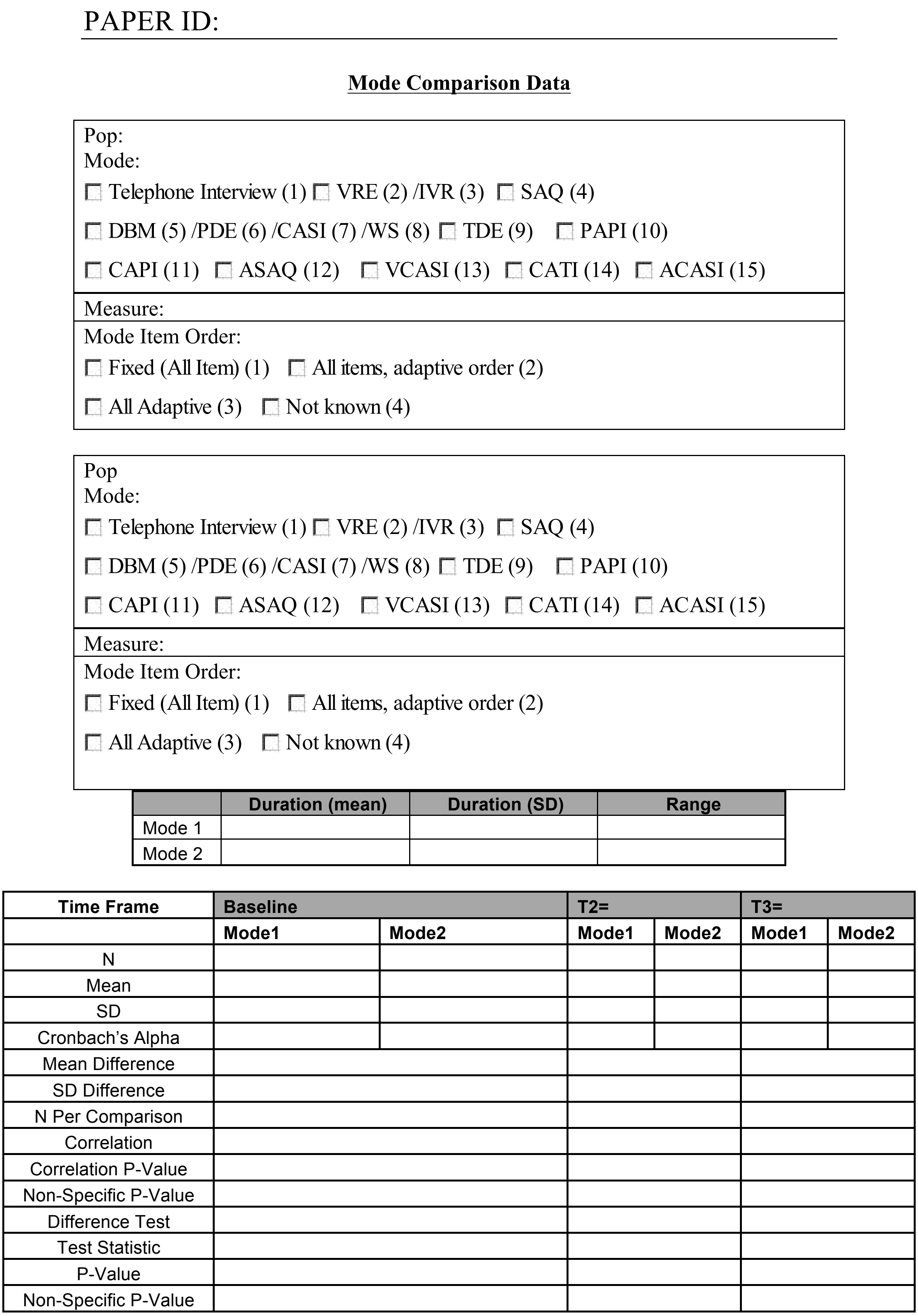
Number of Subjective Items:

Skip Instruction:

□ Yes (1) □ No (0) □ Don't Know (2)

Notes:

Datasheet 6: mode comparison data



PAPER ID:

Mode Comparison Data

Pop:
Mode:
□ Telephone Interview (1) □ VRE (2) /IVR (3) □ SAQ (4)
□ DBM (5) /PDE (6) /CASI (7) /WS (8) □ TDE (9) □ PAPI (10)
□ CAPI (11) □ ASAQ (12) □ VCASI (13) □ CATI (14) □ ACASI (15)

Measure:

Mode Item Order:
□ Fixed (All Item) (1) □ All items, adaptive order (2)
□ All Adaptive (3) □ Not known (4)

Pop
Mode:
□ Telephone Interview (1) □ VRE (2) /IVR (3) □ SAQ (4)
□ DBM (5) /PDE (6) /CASI (7) /WS (8) □ TDE (9) □ PAPI (10)
□ CAPI (11) □ ASAQ (12) □ VCASI (13) □ CATI (14) □ ACASI (15)

Measure:

Mode Item Order:
□ Fixed (All Item) (1) □ All items, adaptive order (2)
□ All Adaptive (3) □ Not known (4)

	Duration (mean)	Duration (SD)	Range
Mode 1			
Mode 2			

Time Frame	Baseline		T2=		T3=	
	Mode1	Mode2	Mode1	Mode2	Mode1	Mode2
N						
Mean						
SD						
Cronbach's Alpha						
Mean Difference						
SD Difference						
N Per Comparison						
Correlation						
Correlation P-Value						
Non-Specific P-Value						
Difference Test						
Test Statistic						
P-Value						
Non-Specific P-Value						

PAPER ID:

Correlation Type:

□ Pearson's (1) □ Spearman's (2) □ Kendal (3) □ Limits of Agreement (4)

□ ICC (5)

Any Mode Related Differences:

□ Yes (1) □ No (Identical) (0)

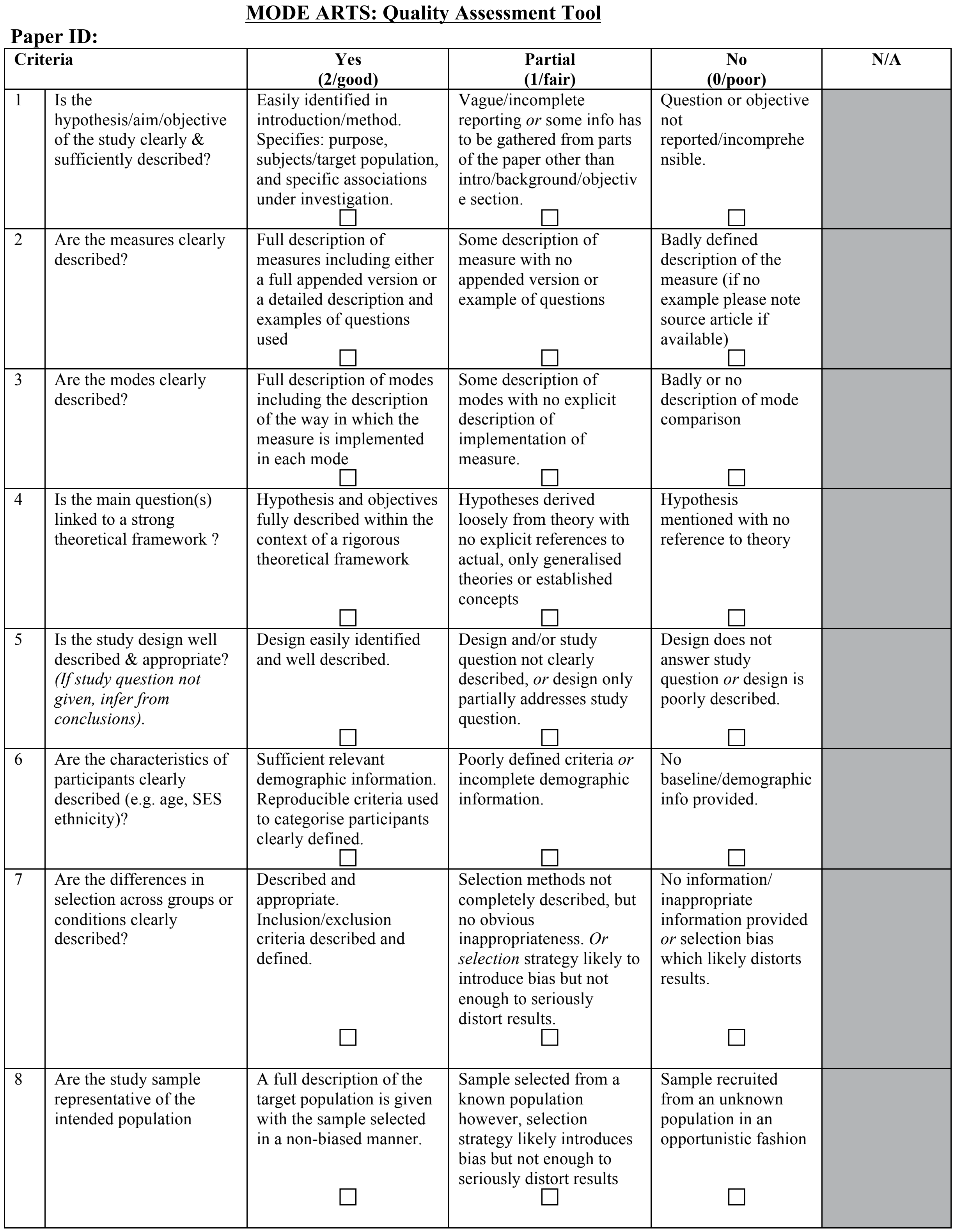
Mode Related Differences Details:

Notes:

	B1	B 1/2	B 0	
A				
A 1/2				
A 0				

Notes:

Datasheet 7: quality assessment



MODE ARTS: Quality Assessment Tool

Paper ID:

Criteria		**Yes (2/good)**	**Partial (1/fair)**	**No (0/poor)**	**N/A**
1	Is the hypothesis/aim/objective of the study clearly & sufficiently described?	Easily identified in introduction/method. Specifies: purpose, subjects/target population, and specific associations under investigation. □	Vague/incomplete reporting *or* some info has to be gathered from parts of the paper other than intro/background/objective section. □	Question or objective not reported/incomprehensible. □	
2	Are the measures clearly described?	Full description of measures including either a full appended version or a detailed description and examples of questions used □	Some description of measure with no appended version or example of questions □	Badly defined description of the measure (if no example please note source article if available) □	
3	Are the modes clearly described?	Full description of modes including the description of the way in which the measure is implemented in each mode □	Some description of modes with no explicit description of implementation of measure. □	Badly or no description of mode comparison □	
4	Is the main question(s) linked to a strong theoretical framework ?	Hypothesis and objectives fully described within the context of a rigorous theoretical framework □	Hypotheses derived loosely from theory with no explicit references to actual, only generalised theories or established concepts □	Hypothesis mentioned with no reference to theory □	
5	Is the study design well described & appropriate? *(If study question not given, infer from conclusions).*	Design easily identified and well described. □	Design and/or study question not clearly described, *or* design only partially addresses study question. □	Design does not answer study question *or* design is poorly described. □	
6	Are the characteristics of participants clearly described (e.g. age, SES ethnicity)?	Sufficient relevant demographic information. Reproducible criteria used to categorise participants clearly defined. □	Poorly defined criteria *or* incomplete demographic information. □	No baseline/demographic info provided. □	
7	Are the differences in selection across groups or conditions clearly described?	Described and appropriate. Inclusion/exclusion criteria described and defined. □	Selection methods not completely described, but no obvious inappropriateness. *Or selection* strategy likely to introduce bias but not enough to seriously distort results. □	No information/ inappropriate information provided *or* selection bias which likely distorts results. □	
8	Are the study sample representative of the intended population	A full description of the target population is given with the sample selected in a non-biased manner. □	Sample selected from a known population however, selection strategy likely introduces bias but not enough to seriously distort results □	Sample recruited from an unknown population in an opportunistic fashion □	

9	How were participants allocated to conditions?	*If randomisation appropriate:* Evidence of well randomised design with a description of the method used (e.g. random number tables, block design). □	No randomisation mentioned but a stratified sampling method is utilised (i.e. may be that full randomisation may not be possible). □	Random allocation not mentioned although it would have been feasible and appropriate (and possible done). □	Study has no control group i.e. observation-al /surveys/ case-control. *Or* adequate justification for non-randomisation given. □
10	Are population characteristics (if measured & described) controlled for and adequately described?	Appropriate control at design/analysis stage *or* randomised study with comparable baseline characteristics. □	Incomplete control/ description. *Or* not considered but unlikely to seriously influence results. □	Not controlled for and likely to seriously influence results. □	
11	Was consideration given for data collected at different times (within groups)	A well described hypothetical reason why data was collected from participants at different time points or comparison with matched historical data set □	Data was collected at different times due to specific opportunity □	No explanation for data collection at different time points, either by chance □	Studies which data was collected at the same time point or between groups □
12	Are the groups adequately compared across	The same measure or mode adapted measures are applied to both groups with full description of procedure □	No clear description of comparison across responder groups only that the same measure was utilised □	No description of methods of comparison between groups or measure application □	Studies that compare different modes within the same group □
13	Have the characteristics of non-responders or participants lost to follow-up been described?	Losses adequately reported & not likely to affect results, *Or* no responders or participants lost to follow up □	Losses not well reported, but small & not likely to affect results. □	No information *or* large losses of responders and likely to affect results. □	
14	Are the main findings clearly described?	Simple outcome data (e.g. mean/proportions) reported for all major findings. □	Incomplete or inappropriate descriptive statistics. □	No/inadequate descriptive statistics □.	
15	Are methods of analysis adequately described and appropriate?	Described and appropriate. □	Not reported but probably appropriate *or* some tests appropriate, some not. □	Methods not described and cannot be determined. □	
16	Are estimates of variance reported for the main results?	Appropriate estimates provided (SD/SE, confidence intervals). □	Undefined *or* estimates provided for some but not all outcomes. □	No information. □	
17	Does the explanation of the results lie within the theoretical framework identified in the introduction	Clear and coherent description of results discussed in relation to previous established theoretical framework □	Findings related to generalised theory with no specific relation to specific theory □	Findings discussed with no consideration to previously mentioned theory □	
18	Are the conclusions supported by the results?	All conclusions supported by data. □	Some of the major conclusions are supported by the data; some are not. *Or* speculative interpretations are not indicated as such. □	None/few of major conclusions supported by the data. □	

Appendix 3

Original funding proposal

Aim

To identify generalisable factors affecting responses to different modes of data collection from a systematic review of the literature.

Objectives

- To review all studies comparing two modes of administration for subjective outcomes and assess the impact of mode of administration on response.
- To provide an overview of the theoretical models of survey response and how they relate to health sciences research.
- To explore the impact of findings for key identified health-related measures.
- To create an accessible resource for health science researchers which will advise on the impact of the selection of different modes of data collection on response.

Outputs

- Generalisable guidance as to differences in the nature of response between modes of data collection.
- Overview of the theory of survey response in relation to measures used in health sciences research.
- Online resource for researchers designing studies.
- Provide workshops to relevant audiences to disseminate the results.

Background

Many studies in health sciences research rely on subjective outcome measures of some form or another. The increasing recognition of the importance of subject attitude to, and perceptions of, health and services provision has led to a rapid growth in such measures. Few clinical trials, even with interventions pharmacological or surgical in nature, would be run today without measuring patients' QoL and the assessment of the acceptability of the intervention being trialled. Survey methodologies (in, for example, the business, marketing, social and political sciences) have an entire literature of their own, covering theory to practice, much of which has been slow to be recognised in the health arena. Few health-related outcome development papers indicate a theoretical approach to eliciting survey response.

Survey response and mode of data collection: psychological theories and survey techniques

The lack of an accepted theoretical basis for survey response was highlighted by Albaum,[141] who noted the distinction between survey techniques and underlying psychological models. Four general theoretical frameworks considered to be particularly relevant to marketing research were reviewed; social exchange theory, cognitive dissonance theory, self-perception theory and theories of commitment and involvement. Albaum *et al.*[141] surveyed the awareness and application of these theoretical models among business researchers across the world and found greatest adherence to theories of commitment and involvement. However, this theoretical review focused upon response decision rather than data quality or nature, although they did comment on the relative application of different models across varying data collection modalities.

Survey non-response and increasing concerns about maintaining adequate levels of response have led researchers to seek to categorise different forms of non-response. For example, Groves and Couper distinguish non-response due to non-contact, refusal to cooperate and inability to participate.[4] The use of incentives to maintain response has, in turn, fostered theoretical development about how such inducements work which, for example, have focused upon economic theories of incentives through to models describing a broader consideration of social exchange. Comprehensive theories of survey involvement have also been introduced and tested empirically[5] (for example Groves *et al.*[18]).

Cognitive approaches to surveying

More recently, a paradigm shift has been described within survey methodology from a statistical model focused upon the consequences of surveying error to social scientific models exploring the causes of error.[6] Attempts to develop such theories of (a) survey error, (b) decisions to participate and (c) response construction have been brought under the general banner of the Cognitive Aspects of Survey Methodology (CASM) movement. Understanding and reducing measurement error, rather than sampling error is at the forefront of this endeavour but Tourangeau notes how the statistical and social scientific approaches are complimentary rather than mutually exclusive. The impetus for recent theoretical developments is very much provided by technological innovation and diversity and a requirement to understand the relative impact of different data collection modes upon survey response.

Several information processing models describing how respondents answer questions have been proposed which share a common core of four basic stages: comprehension of the question; retrieval of information from autobiographical memory; use of heuristic and decision processes to estimate an answer; and response formulation.[7] These models describe mostly sequential processing, apart from that proposed by Willis.[142] The models have contributed to efforts to identify and resolve cognitive response problems in self-report questionnaires and thereby improve data quality in surveys through the use of evaluative and experimental techniques. Examples of the former include cognitive respondent interviews. The potential application of cognitive models and evaluative techniques to subjective self-report in areas such as HRQoL has recently been encouraged.[143,144]

A good example of a sequential information processing model is provided by Tourangeau *et al.*[8] Their model encompasses the four stages described above: (a) comprehension of the survey item; (b) retrieval of relevant information; (c) utilisation of information in making a judgement; and (d) formulating a response. For each stage, there are associated processes identified, which a respondent may or not use when answering a question. Each stage and each process may be a source of response error. The theory is proposed for examining and understanding response to questions about events and behaviour as well as inherently subjective states such as attitudes.

The increase in options for survey data collection

As indicated above, there has been a substantial expansion in the modes of data elicitation and collection available to survey researchers over the last 30 years. In 1998, Tourangeau and Smith[9] identified six methods that may be employed for *face-to-face interviews* including paper-and-pencil personal interviews (PAPIs), paper-and-pencil self-administered questionnaires (SAQs), Walkman-administered questionnaires (audio-SAQs), computer-assisted personal interviews (CAPIs), computer-assisted self-administered interviews (CASIs) and audio computer-assisted self-administered interviews (ACASIs). Subsequently, Tourangeau *et al.*[8] delineated 13 different modes of survey data collection (including remote data collection methods such as telephone, mail, e-mail and the internet), which they considered differed in terms of five characteristics: how respondents were contacted; the presentational medium (e.g. paper or electronic); method of administration (via interviewer or self-administered); sensory input channel used; and response mode.

Applying cognitive models to survey response modality

Psychological models of survey response have been applied to the issue of data collection mode. Tourangeau and Smith[9] proposed three characteristics of the data collection mode that may be affecting response; computerisation, whether a survey schedule is self- or interviewer administered, and whether survey items are read by or to the respondent. A fourth characteristic (the use of telephone) was included in a later formulation of this model.[8] Three psychological variables are considered to mediate the impact of data collection mode; degree of privacy permitted (subsequently amended to 'impersonality'), level of cognitive burden imposed and the sense of legitimacy engendered by the approach.[9,145] The model hypothesises the effect of the mediating variables upon levels of reporting, accuracy, reliability and rate of missing data.

The model has still to be systematically evaluated although some evidence is available. For example, an important consideration has proven to be survey item sensitivity which may serve to emphasise differences between data collection modes (e.g. self-administration vs interviewer). Approval from the interviewer would appear to be the salient influence and may lead to either under- or over-reporting of behaviour depending upon its social acceptability. The level of privacy or degree of impersonality afforded by the data collection mode will thus differentially influence the impact of this tendency. While the studies non-systematically reviewed by Tourangeau and Smith[9] involve behavioural self-report (some of which may be externally validated, e.g. alcohol consumption), other non-observable attitudes may be equally susceptible to such influences (e.g. social stereotyping, racial attitudes, etc.).

Variations even within the same mode of data collection further complicate evaluation. For example, Honaker[10] describes computer administered versions of the MMPI, which differ in terms of type of computer being used, different computer–user interfaces with inconsistent item presentation and response formats. Therefore, results from one computerised version of a test cannot be easily generalised to other versions. Other variables that could mediate the effect of different modes of data collection have also been considered, including the overall pace of the interview, the order of survey item processing and role of different mental models employed by respondents. Although the latter in particular is rarely assessed, it has been considered a potentially significant mediator of response behaviour.[8]

Alternative cognitive approaches include work on optimising and satisficing, concepts described as two ends of a continuum of thoroughness of the response process.[39] A respondent may proceed through each cognitive step less diligently when providing a question response or they may omit the middle two steps completely (i.e. retrieval and judgement) – examples of weak and strong satisficing, respectively. In either situation, a variety of decision heuristics may be utilised by the respondent to provide a satisfactory answer. The theory has been used to explain a variety of phenomenon observed in surveys, for example, response order effects (recency and primacy), which emphasise the role of scale design and mode of administration.

Holbrook *et al.*[42] reviewed survey satisficing theory and another hypothetical source of measurement error, social desirability bias across telephone and face-to-face interviews. The probability of satisficing is a function of respondent ability, respondent motivation and task difficulty. Situational factors such as level of non-verbal communication, interview pace and respondent multitasking, which differ between modes interact with respondent disposition to affect response quality. Social desirability bias whereby respondents intentionally misrepresent themselves in their survey responses may differentially affect data collected via different modes. This could stem from differences in social distance, rapport and trust. Holbrook *et al.*[42] found evidence that suggested that telephone interviews increased satisficing and social desirability response bias compared to face-to-face interviews. Also highlighted was the potential interaction of factors such as educational level.

The challenge for health sciences research

As described above, the first characteristic underlying the different modes of data collection considered by Tourangeau was method of contact.[8] Work assessing the impact of an integrated process of respondent approach, consent and data collection has addressed bias due to selective non-ascertainment (i.e. the exclusion of particular subgroups). This may be clearly identifiable subgroups in terms of people without telephones or computers (for telephone or internet approaches), or less clearly identifiable subgroups, i.e. those with lower levels of literacy or the elderly (for paper-based approaches). There is also considerable work on improving response rates and the biases induced by certain subgroups being less likely to consent to take part in a survey.

Furthermore an important question in Health Sciences Research is the use of data collection methodologies within prospective studies, where patients have already been recruited via another approach. This could be within a clinic or other health service setting rather than the survey instrument being the method of approach as well as data collection. Edwards *et al*. have recently reviewed the literature (both health and non-health) to identify randomised trials of methods of improving response rates to postal questionnaires. Another recent review in health-related research[5] has focused on the completeness of data collection and patterns of missing data, as well as response rates.

Indeed guidance is needed not just in terms of which is the 'best' method to use and most appropriate theoretical model of response, but also the possible effects of combining data collected via different modes as there is an increasing need for multimethod follow-up to capture all of the sample of interest. For example, a commonly observed multimethod approach is when a second mode of data collection is used when the first has been unsuccessful (e.g. using telephone interview when there has been no response to a postal approach[11]). Criteria for judging equivalence of two approaches is therefore required. Honaker[10] uses the concepts of *psychometric equivalence* and *experiential equivalence*. The former describes when the two forms produce results with equal mean scores, identical distribution and ranking of scores and agreement in how scores correlate with other variables. The latter deals with how two forms may differ in how they effect the psychometric and non-psychometric components of the response task.

In order to inform health services research, guidance is needed, which quantifies the differences between modes of data collection and indicates which factors are associated with the magnitude of this difference. These could be *contextual based* in terms of where the participant is when the information is completed (e.g. health setting, own home, work), *content based* in terms of questionnaire topic (e.g. attitudes to sexual behaviour) or *population based* (e.g. elderly). Previous work has shown moderate reliability between SAQ and interview on health problems in an elderly population post transurethral resection of the prostate.[146] However, there was a consistent tendency for the SAQ to underestimate a patient's health problems compared with interview. The factors identified by Tourangeau also need to be tested across a wide range of modes and studies.

Defining subjective outcomes in health sciences research

Of particular interest in HSR is the collection of data which cannot be validated objectively. This results in a situation where there is no 'gold standard' with which to compare results to and therefore care needs to be taken as to the presumption of the 'correctness' of responses. This incorporates many types of outcome which are of key interest to health researchers, such as attitudes, intentions to behave and beliefs about illness. This type of outcome can be classified as evaluation-based,[59] where the subjective perspective of the individual is an intrinsic component of the construct being measured. These can be distinguished from performance- and perception-based measures using the following example (from Schwartz and Rapkin):[62]

- *Performance* Timed walk up flight of stairs.
- *Perception* How often do you walk up a flight of stairs?
- *Evaluation* How difficult is it to walk up a flight of stairs?

The involvement of proxy raters in the assessment process for certain groups, particularly in health, is relatively common. For certain patient groups self-report may be difficult and another person is chosen to report on their behalf. All of the modes and much of the theoretical basis of response described above can be used to collect data about an individual via a proxy. This proxy may be a relative (such as a parent or spouse) or someone responding in a professional capacity, such as a health professional. The focus on an evaluation-based framework for outcome measures would lead to this type of measure being included when the comparison is of different methods of data collection within an individual (i.e. incorporating the same individual's subjective perspective). However, the subjective nature of evaluation-based outcomes which involve judgement using idiosyncratic criteria would lead to studies that compare proxy-reporting to self-reporting being excluded from this review.

Review: direct comparisons of data collection modes (health and non-health-related outcomes)

Methodology

Overview: an extensive search of both health and non-health literature will be conducted to identify studies which compare two or more modes of data collection on subjective measurement on the same scale.

Outcomes measures: evaluation-based measures such as attitude, satisfaction, belief, intention to behave, QoL constructs such as anxiety, pain, vitality (not physical functioning).

Studies: will need to have compared two or more modes of data collection in terms of the responses given. Studies purely considering response rates, data recording errors or costs will not be included.[147] These studies will be identified by the search strategy and the reviews to date have limited themselves to postal[148] and other self-completed surveys.[2] Although it is not covered by this application due to cost limitations and is not specific to the remit of brief, this gives the opportunity to provide a database that can be analysed separately. Response rates and costs of using proxy raters and the impact of the use of information technology in either interviewer assisted or self-completed modes are valid questions still to be answered.

Topics: studies in any topic area, both health and non-health, will be included.

There is considerable literature on the impact of different response options on outcome, therefore this review will be restricted to studies where the sole purpose of a different response scale is to accommodate the data collection mode. An example of this would be where a postal questionnaire uses a visual analogue scale, whilst a telephone interview would have to replace this with an ordinal one. This will be controlled for in the analysis.

Search strategy

McColl *et al.*[5] started their review in 1975 with the justification that this was the decade in which several seminal works on surveys were published and the interest in survey methodology took off. However, Edwards' review[146] on response rates identified a number of randomised trials of methods of increasing response rates to postal surveys published prior to this. Therefore, we intend to search electronic databases from the dates they are available. *Box 1* gives the electronic databases used in previous systematic reviews of response rates (Edwards) and design issues (McColl) plus additional databases felt to be of relevance.

In addition to the above databases, the National Research register will be searched for ongoing relevant studies. Certain non-indexed highly relevant collections will be hand searched (e.g. the proceedings of the Survey Research Methods Section of the American Statistical Association). All included papers will have their reference lists searched for relevant papers, using a pearl-growing approach.

Negative publication bias is unlikely to be operating for this type of study, i.e. whether two methods are shown to be the same or different is unlikely to affect the chances of a study being published. Therefore, the search will be limited to published studies, so databases covering grey literature such as Index to Theses, Dissertation Abstracts and SIGLE will not be searched unless there is a lack of evidence for any particular mode of data collection.

The search strategy will focus on data collection mode with additional filters for identifying comparative studies. A matrix approach will be used to reduce the number of studies that report data using only a single mode of administration which are identified. This approach means that we will be searching for studies which contain any *two* of the following sets of terms:

- Question$or paper or postal or mail
- Telephon$
- Computer$
- Interview$

A scoping search on MEDLINE from 1996 to 2004 gives the following number of hits (*Table 1*).

BOX 1 Electronic databases for searching

Applied Social Science Index and Abstracts (ASSIA)

British Nursing Index (BNI)

Cambridge Scientific Abstracts

Cinahl[a]

Cochrane Controlled Trials Register[a]

EconLit[a]

Educational Resources Information Centre (ERIC)[a]

EMBASE[a]

HMIC (King's Fund and DH Data)

ISI Science Citation Index (SCI)[a]

ISI Social Science Citation Index (SSCI)[a]

MEDLINE[a,b]

PsycINFO[*,a,b]

Social Psychological Educational Criminological Trials Register[a]

Social Service Abstracts[a]

Sociological Abstracts[a]

[a] Used by Edwards *et al.* *Both previous reviews used one of the database within PsycINFO (PsychLIT).
[b] Used by McColl *et al.*

TABLE 1 Hits from initial scoping search of MEDLINE (1996–2004)

	Question$	Telephon$	Computer$	Interview$
Question$	–	4805	12,634	20,361
Telephon$		–	1031	4869
Computer$			–	1462
Interview$				–

These will be combined with appropriate words for each database to focus on studies of reporting validation. In MEDLINE the Mesh term 'Reproducibility of Results' will be used. This reduces the number of hits in the above scope and produces a far more sensitive search (*Table 2*).

All searches will be limited to studies of humans. Non-English studies will be identified but only included where there is a lack of evidence in the comparison of any two particular modes.

All identified titles and abstracts will be downloaded into a Reference Manager database, duplicates removed and then titles and abstracts independently reviewed by two reviewers to assess eligibility for inclusion. Studies which either or both reviewers consider eligible will be retrieved in full. An assessment of chance corrected agreement (Kappa) will be made after every 100 abstracts reviewed as a form of quality control on the process. Full papers will again be reviewed for eligibility and data extracted by two reviewers. Additional searches will be made for the ten authors with the highest number of hits.

TABLE 2 Hits from scoping search of MEDLINE (1996–2004) filtered for 'Reproducibility of Results'

	Question$	Telephon$	Computer$	Interview$
Question$	–	243	921	1477
Telephon$		–	42	258
Computer$			–	105
Interview$				–

Eligibility criteria

Include studies:

- using two or more different modes of data collection used
- measuring an evaluation-based assessment
- where both modes of data collection are applied to the same measurement scale
- that have a comparative element either at an individual or group level.

Exclude studies:

- comparing proxy to self-completion
- focusing solely on response or error rates and cost of administration.

Level of comparative analysis

Individual level comparative studies will consist of individuals being exposed to both modes of data collection and their results being compared in a paired analysis. The highest level of evidence would come from those that randomised each individual as to the order in which the data collection modes were used. This type of study design is essentially a cross over trial, allowing for assessment of carry-over effect (recall bias). Additional quality criteria would be consideration and justification of the impact of the time lapse between approaches and the impact of participant recall and stability of the construct being measured.

Group level comparative studies would involve individuals being randomised (or quasi-randomised) to one of the modes of collection to be compared. Analysis would then be at a group level. Consideration would need to be given within the study to the level of balance that the randomisation/quasi-randomisation had achieved on other factors associated with the outcome.

Data extraction

A data extraction sheet will be developed and piloted covering standard quality markers for the reporting of studies, along with factors specific to individual and group level comparisons. A training set of papers ($n = 25$) will be critically appraised and data extracted by two researchers. After this each paper will have data extracted by a single researcher except where difficulties arise. The extraction of statistical data and statistical modelling will be guided by Dr Hood and Prof. Russell.

Data could be reported in one of the following ways:

- means/mean differences (or proportions) with SE (e.g. Krysan *et al.*[149])

- percentage agreement or Kappa statistics (e.g. Doll *et al.*[147])
- variances or reliability coefficient.

All studies would be rated on their quality of reporting in terms of response rates/loss to follow-up, details of their follow-up procedures for non-response and patterns of missing data. Additional variables will rate the difference between the two methods being tested in terms of development and validation and the intensity of the follow-up process. This will include whether the design was theoretically based. With different modes of data collection, identical follow-up procedures are unlikely to be appropriate; however, they should be equivalent in terms of intensity. This leads to a measurement of quality of study that is based on the degree of similarity between the two approaches. This will be rated in terms of development/validation and intensity of follow-up (same/moderately different/very different). A quality scoring system based on this will be developed and controlled for in the analysis.

Key data defining how respondents were contacted; the presentational medium (e.g. paper or electronic); method of administration (via interviewer or self-administered); sensory input channel used (audio and/or visual); and response mode (verbal or manual) will be identified for each mode within each comparison. Where possible, the content, context and population will be categorised.

Analysis

Information from the data extraction sheet will be entered into SPSS for preliminary analysis. The studies should provide information on overall means/mean differences (for group/individual studies) and standard errors. These will be analysed using meta-regression to explore differences by mode of data collection and other variables of interest such as context, content and population. The dependent variable in these analyses will be the standardised difference between the two modes. The modes will be labelled according to the categories identified from theoretical cognitive models.[8] This would involve modes of data collection categorised according to differences in the presentational medium, method of administration, sensory input and response mode. The impact of levels of computerisation will also be assessed. Where more than one outcome per study is of interest, a two-level model will be fitted (using MlWin) to allow for correlations between outcomes within a study. Assessment will be made whether a fixed or random effect fits best for each factor.

Possible moderating factors will be assessed, covering:

- Administration factors, such as intensity of follow-up.
- Population factors, such as age, social class, educational level and disease group.
- Scale-specific factors, such as number of items, response options, time taken and the theoretical basis for its development. A key variable to explore will be whether the scale is health related or not.

Certain modes of data collection may not be represented in enough of the identified studies to be included in the analysis of moderating effects. This analysis will enable us to ascertain the degree to which generalisable conclusions can be drawn across topic areas and populations.

Other factors that will be explored, provided enough studies are identified, are whether the magnitude of the differences between modes is affected by the number of items to be completed and the time taken. Certainly the degree of recall bias in individual studies may be affected by the number of items being completed.

Individual and group studies will be analysed separately. Sensitivity analysis will explore the impact on the conclusions drawn of weighting the regression by quality and sample size.

In order to show psychometric equivalence it is not enough for the mean differences to be close to zero, the distributional properties must also be the same. Therefore where possible comparison of the variances for the different modes of data collection will also be analysed. Again, group and individual studies will need to be analysed separately. This analysis will use the ratio of the variances for each mode from a study and explore whether particular modes of data collection lead to greater variability in response.

Bringing the results into the health domain

A key question in health sciences research is how generalisable the lessons learnt in other disciplines such as sociology and psychology are to the health field. Certain subjective constructs of interest in these other disciplines are more clearly related to outcomes we wish to measure, although whether the cognitive processes involved in responding are content specific remains to be shown. Therefore we propose to undertake two additional pieces of work.

Review of theory

Much of the theory of survey response is published in the survey methodology literature. Since an essential part of understanding the results from the review is to link it to theory, we will undertake a review of the psychological models which can explain/predict individual response to differing modes of data collection. This will be drawn together and interpreted for the health domain. This will be used to provide guidance for researchers developing new measures – for example, on particular validation assessments needed for different modes of data collection. More generally, it can also help guide good practice in the development, design and application of health outcome measures.

Additional review/overview

The systematic review above is limited to studies which have directly compared different modes of data collection. However, there is still a question whether this type of study generalises to those using a single mode of data collection. The focus on comparing modes of administration may make the studies 'idealised' to certain degree with the typical focus being on recruitment, retention and compliance issues rather than on the construct being measured per se. The presence of participant recall bias may under estimate differences between modes. There is therefore a value in considering whether similar patterns of differences exist in studies which use a single mode of administration to the comparative studies included in the review above.

Only a small number of studies within the health field have directly compared two or more modes of data collection. In contrast, a very large number of studies have each used a single mode of data collection. These single-mode studies can be used to assess the generalisability of the results of the review. The review will identify direct comparisons of generic health-related measures such as SF-36[104,113,114] and condition specific measures.[150]

In order to address this issue of external validity, we will review studies that have used one generic instrument, the SF-36, and up to three condition specific instruments identified during the search for direct comparisons. For these outcomes studies will be identified which have administered (singly) the modes of administration which were directly compared. For the generic measure (SF-36) most of the alternatives (telephone, interview) will have been compared to paper-based questionnaires. In this case we will identify a sample of paper-based studies in the same patient grouping as other modes have been used in. These will be analysed to explore whether the magnitude of the differences between the measures (controlling for differences in study design and population) shown in the direct comparison studies is born through to studies using an individual mode.

Outputs

The results of this systematic review will show the magnitude of differences between different modes of data collection and how this is affected by moderator variables such as context and population. It will also explore further the theoretical framework proposed by Tourangeau. Practical outputs could take the form of actual correction factors for modes with have been well studied (and factors are shown to be generalisable) and more general guidance for the less well studies ones.

The results of the review will be evaluated alongside the cognitive models proposed in the theory of survey response. A particular focus will be on how the models related to measures used in health. The additional review of individual measures will related the findings of from the general review directly to the types of measures of interest in health.

Dissemination

A key component of the dissemination strategy is to provide an online resource for health services researchers. This will include a database summarising all of the direct comparison studies so that they can be easily searched and identified. However, a key component of this will be where possible to indicate quantitatively how different modes will impact on the results during the planning stage of a study. It is also hoped that this initial resource will be contributed to by research teams using novel modes of data collection to provide an ongoing resource which is both used by and contributed to by the whole research community. The ongoing maintenance of such a resource would become part of the Centre for Health Science Research.

In addition to this a workshop will be held to discuss the theoretical perspectives in survey response and how they relate to health. We will also look to target workshops at key conferences such as the International Society for Quality of Life Research (ISOQOL) and the Society for Social Medicine (SSM). In addition to this we will also offer workshops/seminars to key organisations such as the Royal Colleges and the Royal Statistical Society. The components of the review will be written up into a report and as peer reviewed publications in mainstream journals.

Management structure

The co-applicants will form a management team which will meet monthly by audio conference and face-to-face once a quarter, starting with a face-to-face meeting. Members of this team have worked across Wales (and the rest of the UK) using this combination of audio and face-to-face meetings in the past successfully. Dr Kerry Hood, Mike Robling, Lesley Sander and the RA will meet formally on a weekly basis between management meetings. Dr Kerry Hood will lead the meetings, manage the project day to day and have line management responsibility for the two employed staff. The review of theoretical models will be managed by Mike Robling and Dr David Ingledew. Prof Ian Russell will work with Dr Hood on the statistical modelling.

Justification of resources and time frame

This systematic review is across numerous electronic databases and from scoping searches is likely to identify a large number of studies. Therefore it is proposed to employ two members of staff for a year each. Lesley Sander is an experienced information scientist who is available to start immediately. She will initiate the project whilst we are appointing the RA. Resources are requested for PCs for both of these members of staff and an additional copy of Reference Manager (plus manual) for the RA. We estimate needing £2000 for inter-library loans. This cost has been kept down due to the fact that Cardiff University has extensive libraries and e-journals and therefore an amount is requested for photocopying and printing. The Proceedings of the Survey Methods Section of the American Statistical Association are available on CD which has been costed in along with the manual and the electronic bibliography for SF-36. Consultancy time of 12 days for input on the review of theory by Dr David Ingledew (@£500 per day) and 4 days for statistical modelling input from Prof. Ian Russell (@£1000 per day) has been costed.

Costs for travel and telephone for management meetings has been included for six face-to-face meetings (one initial followed by once a quarter) and three audio conferences per quarter.

We are planning the development and design of the web site with a company who undertake much work in the academic health field (waters-design) and have recently developed the website for the new Swansea Clinical School. An approximate costing for this has been put at £8000. In order to ensure dissemination via workshops, a conference budget of £3500 has been requested.

Timetable

Month	Tasks
0–3	Refine search strategy
	Draught data extraction sheet
	Appoint RA
	Run searches and remove duplicates
3–6	Assess abstracts for inclusion
	Retrieve full papers and assess for inclusion
	Pilot data extraction sheet
	Identify specific health-related scales for single-mode studies and search
6–9	Extract data from included papers
	Retrieve papers on single mode
	Identify and retrieve theoretical papers
9–12	Enter extracted data
	Analyse direct comparisons
	Extract data for single-mode studies
	Synthesise the theoretical papers
12–15	Analyse single-mode studies
	Write up report and papers
	Design web page
	Conduct workshops

Appendix 4

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses checklist

Section/topic	#	Checklist item	Reported on page #
Title			
Title	1	Identify the report as a systematic review, meta-analysis or both	i
Abstract			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants; and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number	xi–xv
Introduction			
Rationale	3	Describe the rationale for the review in the context of what is already known	1–3
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes and study design (PICOS)	3
Methods			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g. web address), and, if available, provide registration information including registration number	NA – original funding proposal pp. 103–114
Eligibility criteria	6	Specify study characteristics (e.g. PICOS, length of follow-up) and report characteristics (e.g. years considered, language, publication status) used as criteria for eligibility, giving rationale	19–20
Information sources	7	Describe all information sources (e.g. databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched	85
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated	85–7
Study selection	9	State the process for selecting studies (i.e. screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis)	22–3
Data collection process	10	Describe method of data extraction from reports (e.g. piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators	24
Data items	11	List and define all variables for which data were sought (e.g. PICOS, funding sources) and any assumptions and simplifications made	95–102
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis	25
Summary measures	13	State the principal summary measures (e.g. risk ratio, difference in means)	26
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g. I^2) for each meta-analysis	26–7
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g. publication bias, selective reporting within studies)	26
Additional analyses	16	Describe methods of additional analyses (e.g. sensitivity or subgroup analyses, meta-regression), if done, indicating which were prespecified	27

Section/topic	#	Checklist item	Reported on page #
Results			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram	29
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g. study size, PICOS, follow-up period) and provide the citations	Summary presented given number of studies: pp. 30–2
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12)	Summary presented given number of studies: pp. 32–3
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and CIs, ideally with a forest plot	Only done for SF-36 and MMPI analyses due to number of studies: pp. 42–66
Synthesis of results	21	Present results of each meta-analysis done, including CIs and measures of consistency	36–69
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see item 15)	NA
Additional analysis	23	Give results of additional analyses, if done [e.g. sensitivity or subgroup analyses, meta-regression (see item 16)]	36–69
Discussion			
Summary of evidence	24	Summarise the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g. health-care providers, users and policy-makers)	67–9
Limitations	25	Discuss limitations at study and outcome level (e.g. risk of bias) and at review level (e.g. incomplete retrieval of identified research, reporting bias)	68–9
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research	69
Funding			
Funding	27	Describe sources of funding for the systematic review and other support (e.g. supply of data); role of funders for the systematic review	HTA review

HTA, health technology assessment; NA, not applicable.

Appendix 5

Included papers

Abdoh A, Krousel-Wood MA, Re RN. Validity and reliability assessment of an automated telephone survey system (208). *Congress of Epidemiology Abstracts* 2001:S87.

Addington-Hall J, Walker L, Jones C, Karlsen S, McCarthy M. A randomised controlled trial of postal versus interviewer administration of a questionnaire measuring satisfaction with, and use of, services received in the year before death. *J Epidemiol Community Health* 1998;**52**:802–7.

Aertgeerts B, Buntinx F, Fevery J, Ansoms S. Is there a difference between CAGE interviews and written CAGE questionnaires? ACER 2000;**24**:733–6.

Agel J, Rockwood T, Mundt JC, Greist JH, Swiontkowski M. Comparison of interactive voice response and written self-administered patient surveys for clinical research. *Orthopedics* 2001;**24**:1155–7.

Allison T, Ahmad T, Brammah T, Symmons D, Urwin M. Can findings from postal questionnaires be combined with interview results to improve the response rate among ethnic minority populations? Ethn Health 2003;**8**:63–9.

Amodei N, Katerndahl DA, Larme AC, Palmer R. Interview versus self-answer methods of assessing health and emotional functioning in primary care patients. *Psychol Rep* 2003;**92**:937–48.

Andersson G, Kaldo-Sandstrom V, Strom L, Stromgren T. Internet administration of the Hospital Anxiety and Depression Scale in a sample of tinnitus patients. *J Psychosom Res* 2003;**55**:259–62.

Andersson G, Lindvall N, Hursti T, Carlbring P. Hypersensitivity to sound (hyperacusis): a prevalence study conducted via the Internet and post. *Int J Audiol* 2002;**41**:545–54.

Aneshensel CS, Frerichs RR, Clark VA, Yokopenic PA. Measuring depression in the community: a comparison of telephone and personal interviews. *Pub Opin Q* 1982;**46**:110–21.

Aneshensel CS, Frerichs RR, Clark VA, Yokopenic PA. Telephone versus in-person surveys of community health status. *Am J Public Health* 1982;**72**:1017–21.

Armstrong CS, Sun Z, David TE. Follow up of patients after valvular surgery: mail vs. telephone. *J Heart Valve Dis* 1995;**4**:346–9.

Athale N, Sturley A, Skoczen S, Kavanaugh A, Lenert L. A web-compatible instrument for measuring self-reported disease activity in arthritis. *J Rheumatol* 2004;**31**:223–8.

Augustine AJS. The use of the telephone interview in obtaining information of a sensitive nature: a comparative study. *J Am Stat Assoc* 1978:559–61.

Baer L, Brown-Beasley MW, Sorce J, Henriques AI. Computer-assisted telephone administration of a structured interview for obsessive-compulsive disorder. *A J Psychiatr* 1993;**150**:1737–8.

Bagley C, Genuis M. Psychology of computer use: XX. Sexual abuse recalled: evaluation of a computerized questionnaire in a population of young adult males. *Percept Mot Skills* 1991;**72**:287–8.

Bajos N, Spira A, Ducot B, Messiah A. Analysis of sexual behaviour in France (ACSF): A comparison between two modes of investigation: telephone survey and face-to-face survey. *AIDS* 1992;**6**:315–23.

Ballard C, Prine R. Citizen perceptions of community policing: comparing internet and mail survey responses. *Soc Sci Comput Rev* 2002;**20**:485–93.

Bandilla W, Bosnjak M, Altdorfer P. Survey administration effects? A Comparison of web-based and traditional written self-administered surveys using the ISSP Environment Module. *Soc Sci Comput Rev* 2003;**21**:235–43.

Barry MJ, Fowler FJ, Chang Y, Liss CL, Wilson H, M. Stek, Jr. The American Urological Association symptom index: does mode of administration affect its psychometric properties? *J Urol* 1995;**154**:1056–9.

Bartram D, Brown A. Information exchange article: online testing: mode of administration and the stability of OPQ 32i scores. *Int J Select Assess* 2004;**12**:278–84.

Bauer M, Bohrer H, Aichele G, Bach A, Martin E. Measuring patient satisfaction with anaesthesia: perioperative questionnaire versus standardised face-to-face interview. *Acta Anaesthesiol Scand* 2001;**45**:65–72.

Bauer M, Grof P, Gyulai L, Rasgon N, Glenn T, Whybrow PC. Using technology to improve longitudinal studies: self-reporting with ChronoRecord in bipolar disorder. *Bipolar Disord* 2004;**6**:67–74.

Bausell RB, Rinkus AJ. A comparison of written versus oral interviews. *Eval Health Prof* 1979;**2**:477–86.

Bellamy N, Campbell J, Hill J, Band P. A comparative study of telephone versus onsite completion of the WOMAC 3.0 osteoarthritis index. *J Rheumatol* 2002;**29**:783–6.

Bellamy N, Campbell J, Stevens J, Pilch L, Stewart C, Mahmood Z. Validation study of a computerized version of the Western Ontario and McMaster Universities VA3.0 Osteoarthritis Index. *J Rheumatol* 1997;**24**:2413–5.

Bennett SJ, Perkins SM, Lane KA, Forthofer MA, Brater DC, Murray MD. Reliability and validity of the compliance belief scales among patients with heart failure. *Heart Lung* 2001;**30**:177–85.

Bernadt MW, Daniels OJ, Blizard RA, Murray RM. Can a computer reliably elicit an alcohol history? *Br J Addict* 1989;**84**:405–11.

Berthelsen CL, Stilley KR. Automated personal health inventory for dentistry: a pilot study. *J Am Dent Assoc* 2000;**131**:59–66.

Best SJ, Krueger B, Hubbard C, Smith A. An assessment of the generalizability of Internet surveys. *Soc Sci Comput Rev* 2001;**19**:131–45.

Bishop GF, Fisher BS. Secret ballots and self-reports in an exit-poll experiment. *Pub Opin Q* 1995;**59**:568–88.

Biskin BH, Kolotkin RL. Effects of computerized administration on scores on the Minnesota Multiphasic Personality Inventory. *Appl Psychol Meas* 1977;**1**:543–9.

Bjorner JB, Kosinski M, Ware JE, Jr. Using item response theory to calibrate the Headache Impact Test (HIT) to the metric of traditional headache scales. *Qual Life Res* 2003;**12**:981–1002.

Black CM, Wilson GT. Assessment of eating disorders: interview versus questionnaire. *Int J Eat Disord* 1996;**20**:43–50.

Bliven BD, Kaufman SE, Spertus JA. Electronic collection of health-related quality of life data: validity, time benefits, and patient preference. *Qual Life Res* 2001;**10**:15–22.

Bongers IM, Oers JAV. Mode effects on self-reported alcohol use and problem drinking: mail questionnaires and personal interviewing compared. *J Stud Alcohol* 1998;**59**:280–5.

Booth-Kewley S, Edwards JE, Rosenfeld P. Impression management, social desirability, and computer administration of attitude questionnaires: does the computer make a difference? *J Appl Psychol* 1992;**77**:562–6.

Bouman TK, Schaufeli WB. Equivalence and evaluation of conventional and computer administrations of a symptom check list. *Ned Tijdschr Psychol* 1988;**43**:86–92.

Bower P, Macdonald W, Sibbald B, Garralda E, Kramer T, Bradley S, *et al.* Postal survey of services for child and adolescent mental health problems in general practice in England. *Prim Care Ment Health* 2003;**1**:17–26.

Bower P, Roland MO. Bias in patient assessments of general practice: general practice assessment survey scores in surgery and postal responders. *Br J Gen Pract* 2003;**53**:126–8.

Bowling A, Bond M, Jenkinson C, Lamping DL. Short Form 36 (SF-36) Health Survey questionnaire: which normative data should be used? Comparisons between the norms provided by the Omnibus Survey in Britain, the Health Survey for England and the Oxford Healthy Life Survey. *J Publ Health Med* 1999;**21**:255–70.

Bowman MA, Sharp PC, Herndon A, Dignan MB. Methods for determining patient improvement following visits to family physicians. *Family Med* 1990;**22**:275–8.

Boyer KK, Olson JR, Calantone RJ, Jackson EC. Print versus electronic surveys: a comparison of two data collection methodologies. *J Oper Manag* 2002;**20**:357–73.

Boyes A, Newell S, Girgis A. Rapid assessment of psychosocial well-being: are computers the way forward in a clinical setting? *Qual Life Res* 2002;**11**:27–35.

Bozlu M, Doruk E, Akbay E, Ulusoy E, Cayan S, Acar D, *et al.* Effect of administration mode (patient vs physician) and patient's educational level on the Turkish version of the International Prostate Symptom Score. *Int J Urol* 2002;**9**:417–21.

Brambilla DJ, McKinlay SM. A comparison of responses to mailed questionnaires and telephone interviews in a mixed mode health survey. *Am J Epidemiol* 1987;**126**:962–71.

Bredart A, Mignot V, Rousseau A, Dolbeault S, Beauloye N, Adam V, *et al.* Validation of the EORTC QLQ-SAT32 cancer inpatient satisfaction questionnaire by self- versus interview-assessment comparison. *Patient Educ Couns* 2004;**54**:207–12.

Bremer BA, McCauley CR. Quality-of-life measures: hospital interview versus home questionnaire. *Health Psychol* 1986;**5**:171–7.

Bressani RV, Downs A. Youth independent living assessment: testing the equivalence of web and paper/pencil versions of the Ansell-Casey Life Skills Assessment. *Comput Hum Behav* 2002;**18**:453–64.

Brewer NT, Hallman WK, Fiedler N, Kipen HM. Why do people report better health by phone than by mail? *Med Care* 2004;**42**:875–83.

Brogger J, Bakke P, Eide GE, Gulsvik A. Comparison of telephone and postal survey modes on respiratory symptoms and risk factors. *Am J Epidemiol* 2002;**155**:572–6.

Bryson SE, Pilon DJ. Sex differences in depression and the method of administering the Beck Depression Inventory. *J Clin Psychol* 1984;**40**:529–34.

Buchanan T, Smith JL. Using the Internet for psychological research: personality testing on the World Wide Web. *Br J Psychol* 1999;**90**:125–44.

Bulpitt CJ, Fletcher AE. Quality of life in hypertensive patients on different antihypertensive treatments: rationale for methods employed in a multicenter randomized controlled trial. *J Cardiovasc Pharmacol* 1985;**7**:S137–45.

Bungey JB, Pols RG, Mortimer KP, Frank OR, Skinner HA. Screening alcohol & drug use in a general practice unit: comparison of computerised and traditional methods. *Community Health Stud* 1989;**13**:471–83.

Burke WJ, Rangwani S, Roccaforte WH, Wengel SP, Conley DM. The reliability and validity of the collateral source version of the Geriatric Depression Rating Scale administered by telephone. *Int J Geriatr Psychiatr* 1997;**12**:288–94.

Burke WJ, Roccaforte WH, Wengel SP, Conley DM, Potter JF. The reliability and validity of the Geriatric Depression Rating Scale administered by telephone. *J Am Geriatr Soc* 1995;**43**:674–9.

Burroughs TE, Waterman BM, Cira JC, Desikan R, Dunagan WC. Patient satisfaction measurement strategies: a comparison of phone and mail methods. *Joint Comm J Qual Improv* 2001;**27**:349–61.

Bushnell DM, Martin ML, Parasuraman B. Electronic versus paper questionnaires: a further comparison in persons with asthma. *J Asthma* 2003;**40**:751–62.

Butler N, Newton T, Slade P. Validation of a computerized version of the SCANS questionnaire. *Int J Eat Disord* 1989;**8**:239–41.

Calvet X, Bustamante E, Montserrat A, Roque M, Campo R, Gene E, *et al.* Validation of phone interview for follow-up in clinical trials on dyspepsia: evaluation of the Glasgow Dyspepsia Severity Score and a Likert-scale symptoms test. *Eur J Gastroenterol Hepatol* 2000;**12**:949–53.

Cam K, Akman Y, Cicekci B, Senel F, Erol A. Mode of administration of international prostate symptom score in patients with lower urinary tract symptoms: physician vs self. *Prostate Cancer Prostatic Dis* 2004;**7**:41–4.

Campen Cv, Sixma H, Kerssens JJ, Peters L. Comparisons of the costs and quality of patient data collection by mail versus telephone versus in-person interviews. *Eur J Public Health* 1998;**8**:66–70.

Canoune HL, Leyhe EW. Human versus computer interviewing. *J Pers Assess* 1985;**49**:103–6.

Canterino JC, VanHorn LG, Harrigan JT, Ananth CV, Vintzileos AM. Domestic abuse in pregnancy: a comparison of a self-completed domestic abuse questionnaire with a directed interview. *Am J Obstet Gynecol* 1999;**181**:1049–51.

Carini R, Hayek JC, Kuh GD, Kennedy JM, Ouimet JA. College student responses to web and paper surveys: does mode matter? *Res High Educ* 2003;**44**:19 January.

Caro J, Caro I, Caro J, Wouters F, Juniper EF. Does electronic implementation of questionnaires used in asthma alter responses compared to paper implementation? *Qual Life Res* 2001;**10**:683–91.

Carrete P, Augustovski F, Gimpel N, Fernandez S, Paolo RD, Schaffer I, *et al.* Validation of a telephone-administered geriatric depression scale in a hispanic elderly population. See comment. *J Gen Intern Med* 2001;**16**:446–50.

Caserta MS, Lund DA, Dimond MF. Assessing interviewer effects in a longitudinal study of bereaved elderly adults. *J Gerontol* 1985;**40**:637–40.

Catania JA, McDermott LJ, Pollack LM. Questionnaire response bias and face-to-face interview sample bias in sexuality research. *J Sex Res* 1986;**22**:52–72.

Cates WM. A small-scale comparison of the equivalence of paper-and-pencil and computerized versions of student end-of-course evaluations. *Comput Hum Behav* 1993;**9**:401–9.

Chambers LW, Haight M, Norman G, MacDonald L. Sensitivity to change and the effect of mode of administration on health status measurement. *Med Care* 1987;**25**:470–80.

Chan KS, Orlando M, Ghosh-Dastidar B, Duan N, Sherbourne CD. The interview mode effect on the Center for Epidemiological Studies Depression (CES-D) scale: an item response theory analysis. *Med Care* 2004;**42**:281–9.

Chestnutt IG, Morgan MZ, Hoddell C, Playle R. A comparison of a computer-based questionnaire and personal interviews in determining oral health-related behaviours. *Community Dent Oral Epidemiol* 2004;**32**:410–17.

Chwalow A, Balkau B, Costagliola D, Deeds SG. Comparison of different data collection methods within a study sample: telephone versus home interviews. *Health Educ Res* 1989;**4**:321–8.

Cohen DB. Relation of anxiety level and defense style to frequency of dream recall estimated by different methods. *Psychophysiology* 1968:224.

Collins FE, Jones KV. Investigating dissociation online: validation of a web-based version of the dissociative experiences scale. *J Trauma Dissociation* 2004;**5**:133–47.

Cook AJ, Roberts DA, Henderson MD, Winkle LCV, Chastain DC, Hamill-Ruth RJ. Electronic pain questionnaires: a randomized, crossover comparison with paper questionnaires for chronic pain assessment. *Pain* 2004;**110**:310–17.

Cook DJ, Guyatt GH, Juniper E, Griffith L, McIlroy W, Willan A, *et al.* Interviewer versus self-administered questionnaires in developing a disease-specific, health-related quality of life instrument for asthma. *J Clin Epidemiol* 1993;**46**:529–34.

Coon GM, Pena D, Illich PA. Self-efficacy and substance abuse: assessment using a brief phone interview. *J Subst Abuse Treat* 1998;**15**:385–91.

Corman SR. Computerized vs pencil and paper collection of network data. *Soc Networks* 1990;**12**:375–84.

Couper MP, Rowe B. Evaluation of a computer-assisted self interview (CASI) component in CAPI survey. *J Am Stat Assoc* 1995:1017–22.

Coyne KS, Margolis MK, Gilchrist KA, Grandy SP, Hiatt WR, Ratchford A, *et al.* Evaluating effects of method of administration on Walking Impairment Questionnaire. *J Vasc Surg* 2003;**38**:296–304.

Cronk BC, West JL. Personality research on the Internet: a comparison of web-based and traditional instruments in take-home and in-class settings. *Behav Res Methods* 2002;**34**:177–80.

Davis C, Cowles M. Automated psychological testing: method of administration, need for approval, and measures of anxiety. *Educ Psychol Meas* 1989;**49**:311–20.

Davis LJ, Morse RM. Self-administered alcoholism screening test: a comparison of conventional versus computer-administered formats. *ACER* 1991;**15**:155–7.

Davis RN. Web-based administration of a personality questionnaire: comparison with traditional methods. *Behav Res Methods* 1999;**31**:572–7.

Dillman DA, Brown TL, Carlson JE, Carpenter EH. Effects of category order on answers in mail and telephone surveys. *Rural Sociol* 1995;**60**:674–87.

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. *Soc Sci Med* 1991;**33**:1303–8.

Dommeyer CJ, Baum P, Hanna RW, Chapman KS. Gathering faculty teaching evaluations by in-class and online surveys: their effects on response rates and evaluations. *Assess Eval High Educ* 2004;**29**:611–23.

Eaton J, Struthers CW. Using the internet for organizational research: a study of cynicism in the workplace. *Cyberpsychology Behav* 2002;**5**:305–13.

Eisen SV. Assessment of subjective distress by patients' self-report versus structured interview. *Psychol Rep* 1995;**76**:35–9.

Elliott DB, Fowler FJ. Response order effects in the medicare population: the interaction between mode of survey administration and respondent age. *J Am Stat Assoc* 2000:936–40.

Epstein J, Klinkenberg W, Wiley D, McKinley L. Insuring sample equivalence across internet and paper-and-pencil assessments. *Comput Hum Behav* 2001;**17**:339–46.

Escobedo LG, Landen MG, Axtell CD, Kaigh WD. Usefulness of telephone risk factor surveys in the New Mexico border region. *Am J Prev Med* 2002;**23**:22 July.

Etter JF, Perneger TV. A comparison of cigarette smokers recruited through the internet or by mail. *Int J Epidemiol* 2001;**30**:521–5.

Evan WM, Miller JR. Differential effects on response bias of computer vs. conventional administration of a social science questionnaire: an exploratory methodological experiment. *Behav Sci* 1969:216–27.

Evans DC, Garcia DJ, Garcia DM, Baron RS. In the privacy of their own homes: using the internet to assess racial bias. *Pers Soc Psychol Bull* 2003;**29**:273–84.

Evans M, Kessler D, Lewis G, Peters TJ, Sharp D. Assessing mental health in primary care research using standardized scales: can it be carried out over the telephone? Psychol Med 2004;**34**:157–62.

Fan X, Gong Y, Wei Y. Comparing the computer and paper-and-pencil administrations of the Chinese version of EPQ. *CMHJ* 2004;**18**:276–7.

Farnworth M, Bennett K, West VM. Mail vs. telephone surveys of criminal justice attitudes: a comparative analysis. *J Quant Criminol* 1996;**12**:113–33.

Farrell AD, Camplair PS, McCullough L. Identification of target complaints by computer interview: evaluation of the computerized assessment system for psychotherapy evaluation and research. *J Consult Clin Psychol* 1987;**55**:691–700.

Fenig S, Levav I, Kohn R, Yelin N. Telephone vs face-to-face interviewing in a community psychiatric survey. *Am J Public Health* 1993;**83**:896–8.

Finegan JE, Allen NJ. Computerized and written questionnaires: are they equivalent? Comput Hum Behav 1994;**10**:483–96.

Flyger HL, Kallestrup EB, Mortensen SO. Validation of a computer version of the patient-administered Danish prostatic symptom score questionnaire. *Scand J Urol Nephrol* 2001;**35**:196–9.

Ford B, Vitelli R, Stuckless N. The effects of computer versus paper-and-pencil administration on measures of anger and revenge with an inmate population. *Comput Hum Behav* 1996;**12**:159–66.

Fortner B, Okon T, Schwartzberg L, Tauer K, Houts AC. The Cancer Care Monitor: psychometric content evaluation and pilot testing of a computer administered system for symptom screening and quality of life in adult cancer patients. *J Pain Symptom Manag* 2003;**26**:1077–92.

Fouladi RT, McCarthy CJ, Moller NP. Paper-and-pencil or online? Evaluating mode effects on measures of emotional functioning and attachment. *Assessment* 2002;**9**:204–15.

Fowler FJ, Gallagher PM. Mode effects and consumer assessments of health plans. *J Am Stat Assoc* 1997:928–33.

Fowler FJ, Roman AM, Di ZX. Mode effects in a survey of medicare prostate surgery patients. *J Am Stat Assoc* 1993:730–5.

Fowler FJ, Roman AM, Di ZX. Mode effects in a survey of medicare prostate surgery patients. *Pub Opin Q* 1998;**62**:29–46.

Franke GH. Implications of computer administration on the Frieburger Personality Inventory: two experimental studies. *Z Exp Psychol* 1997;**44**:332–56.

Franke GH. Effects of computer administration on the Symptom Checklist (SCL-90-R) with a special focus on the item sequence. *Diagnostica* 1999;**45**:147–53.

French SA, Peterson CB, Story M, Anderson N, Mussell MP, Mitchell JE. Agreement between survey and interview measures of weight control practices in adolescents. *Int J Eat Disord* 1998;**23**:45–56.

Frick U, Rehm J, Thien U, Spuhler T. Construction and validation of an indicator for alcohol problems in the Swiss Health Survey. *Soz Praventivmed* 1996;**41**:133–42.

Frost NA, Sparrow JM, Hopper CD, Peters TJ. Reliability of the VCM1 Questionnaire when administered by post and by telephone. *Ophthalmic Epidemiol* 2001;**8**: 1 November.

Gaertner J, Elsner F, Pollmann-Dahmen K, Radbruch L, Sabatowski R. Electronic pain diary: a randomized crossover study. *J Pain Symptom Manag* 2004;**28**:259–67. [Erratum appears in *J Pain Symptom Manage* 2004;**28**:626.]

Gano-Phillips S, Fincham FD. Assessing marriage via telephone interviews and written questionnaires: a methodological note. *J Marriage Fam* 1992;**54**:630–5.

Ganse Wv, Hoorne Mv, Backer GD, Pannier R, Vuylsteek K. Comparison between the interview and the self-subkitted questionnaire of the Rose in a pilot study of Artheriosclerosis. *Rev Epidemiologie Sante Publique* 1972;**20**:7–14.

Garcia-Losa M, Unda M, Badia X, Rodriguez-Alcantara F, Carballido J, Dal-Re R, *et al.* Effect of mode of administration on I-PSS scores in a large BPH patient population. *Eur Urol* 2001;**40**:451–7.

Gati I, Saka N. Internet-based versus paper-and-pencil assessment: measuring career decision-making difficulties. *JCA* 2001;**9**:397–416.

Gaudron J-P. The effects of computer anxiety on self-description with a computerized personality inventory. *Eur Rev Appl Psychol* 2000;**50**:431–6.

Geissler A, Paoli K, Maitrejean C, Durand-Gasselin J. Rates of potential and actual cornea donation in a general hospital: impact of exhaustive death screening and surrogate phone consent. *Transplant Proc* 2004;**36**:2894–5.

George CE, Lankford J, Wilson SE. The effects of computerized versus paper-and-pencil administration on measures of negative affect. *Comput Hum Behav* 1992;**8**:203–9.

Gervil M, Ulrich V, Olesen J, Russell MB. Screening for migraine in the general population: validation of a simple questionnaire. *Cephalalgia* 1998;**18**:342–8.

Gibson FK, Hawkins BW. Interviews versus questionnaires. *Am Behav Sci* 1968;**12**:NS–11.

Gitzinger I. Acceptance of tests presented on a personal computer by inpatients. *Psychother Psychosom Med Psychol* 1990;**40**:143–5.

Glaze R, Cox JL. Validation of a computerised version of the 10-item (self-rating) Edinburgh Postnatal Depression Scale. *J Affect Disord* 1991;**22**:73–7.

Gold DR, Weiss ST, Tager IB, Segal MR, Speizer FE. Comparison of questionnaire and diary methods in acute childhood respiratory illness surveillance. *Am Rev Respir Dis* 1989;**139**:847–9.

Gomez-Peresmitre G, Granados A, Jauregui J, Garcia GP, Ramos SAT. Body image measurement: paper and pencil and computerized test versions. *Rev Mex Psicol* 2000;**17**:89–99.

Gonzalez GM, Costello CR, Tourette TRL, Joyce LK, Valenzuela M. Bilingual telephone-assisted computerized speech-recognition assessment: is a voice-activated computer program a culturally and linguistically appropriate tool for screening depression in English and Spanish? Cultur Divers Mental Health 1997;**3**:93–111.

Grappey C. Fiabilite des resultats de methode d'evaluation contingente et modes d'interrogation. *Econ Rurale* 1999;**254**:45–53.

Greenberg A, Manfield MN. On the reliability of mail questionnaires in product tests. *J Mark* 1957;**21**:342–5.

Gretes JA, Songer T. Validation of the Learning Style Survey: an interactive videodisc instrument. *Educ Psychol Meas* 1989;**49**:235–41.

Grossarth-Maticek R, Eysenck HJ, Barrett P. Prediction of cancer and coronary heart disease as a function of method of questionnaire administration. *Psychol Rep* 1993;**73**:943–59.

Hagen K, Zwart JA, Vatten L, Stovner LJ, Bovim G. Head-HUNT: validity and reliability of a headache questionnaire in a large population-based study in Norway. *Cephalalgia* 2000;**20**:244–51.

Hajebrahimi S, Corcos J, Lemieux MC. International consultation on incontinence questionnaire short form: comparison of physician versus patient completion and immediate and delayed self-administration. *Urology* 2004;**63**:1076–8.

Hall MF. Patient satisfaction or acquiescence? Comparing mail and telephone survey results. *J Healthc Mark* 1995;**15**:54–61.

Hallen H, Djupesland P, Kramer J, Toll K, Graf P. Evaluation of a new method for assessing symptoms. *J Otorhinolaryngol Relat Spec* 2001;**63**:92–5.

Han C, Lee B-W, Ro K-K. The choice of a survey mode in country image studies. *J Bus Res* 1994;**29**:151–62.

Hanna AW, Pynsent PB, Learmonth DJ, Tubbs ON. A comparison of a new computer-based interview for knee disorders with conventional history taking. *Knee* 1999;**6**:245–56.

Harewood GC, Wiersema MJ, Groen PCd. Utility of web-based assessment of patient satisfaction with endoscopy. *Am J Gastroenterol* 2003;**98**:1016–21.

Harewood GC, Yacavone RF, Locke GR, 3rd, Wiersema MJ. Prospective comparison of endoscopy patient satisfaction surveys: e-mail versus standard mail versus telephone. *Am J Gastroenterol* 2001;**96**:3312–17.

Harrell TH, Lombardo TA. Validation of an automated 16PF administration procedure. *J Pers Assess* 1984;**48**:638–42.

Hart RR, Goldstein MA. Computer-assisted psychological assessment. *Comput Hum Serv* 1985;**1**:69–75.

Havice MJ, Banks MJ. Live and automated telephone surveys: a comparison of human interviewers and an automated technique'. *J Market Res Soc* 1992;**33**:91–102.

Hawthorne G. The effect of different methods of collecting data: mail, telephone and filter data collection issues in utility measurement. *Qual Life Res* 2003;**12**:1081–8.

Hebert R, Bravo G, Korner-Bitensky N, Voyer L. Refusal and information bias associated with postal questionnaires and face-to-face interviews in very elderly subjects. *J Clin Epidemiol* 1996;**49**:373–81.

Helgeson JG, Ursic ML. The decision process equivalency of electronic versus pencil-and-paper data collectionm. *Soc Sci Comput Rev* 1989;**7**:296–310.

Henson R, Cannell CF, Roth AV. Effects of interview mode on reporting of moods, symptoms, and need for social approval. *J Soc Psychol* 1978;**105**:123–9.

Herzog A, Rodgers WL. Interviewing older adults: Mode comparison using data from a face-to-face survey and a telephone resurvey. *Pub Opin Q* 1988;**52**:84–99.

Heuser J, Geissner E. Computerized version of the pain experience scale: a study of equivalence. German. *Schmerz* 1998;**12**:205–8.

Hinkle AL, King GD. A comparison of three survey methods to obtain data for community mental health program planning. *Am J Community Psychol* 1978;**6**:389–97.

Hinkle J, Sampson JP, Radonsky V. Computer-assisted versus paper-and-pencil assessment of personal problems in a clinical population. *Comput Hum Behav* 1991;**7**:237–42.

Hoher J, Bach T, Munster A, Bouillon B, Tiling T. Does the mode of data collection change results in a subjective knee score? Self-administration versus interview. *Am J Sports Med* 1997;**25**:642–7.

Honaker L, Harrell TH, Buffaloe JD. Equivalency of Microtest computer MMPI administration for standard and special scales. *Comput Hum Behav* 1988;**4**:323–37.

Horswill MS, Coster ME. User-controlled photographic animations, photograph-based questions, and questionnaires: three internet-based instruments for measuring drivers' risk-taking behavior. *Behav Res Methods* 2001;**33**:46–58.

Houck PR, Spiegel DA, Shear MK, Rucci P. Reliability of the self-report version of the panic disorder severity scale. *Depress Anxiety* 2002;**15**:183–5.

Hunt DL, Haynes RB, Hayward RS, Pim MA, Horsman J. Automated direct-from-patient information collection for evidence-based diabetes care. Proceedings/AMIA Annual Fall Symposium 1997:81–5.

Hutchison J, Tollefson N, Wigington H. Response bias in college freshmen's responses to mail surveys. *Res High Educ* 1987;**26**:99–106.

Izquierdo-Porrera AM, Manchanda R, Powell CC, Sorkin JD, Bradham DD. Factors influencing the use of computer technology in the collection of clinical data in a predominantly African-American population. *J Am Geriatr Soc* 2002;**50**:1411–15.

Jennings KD, Stagg V, Pallay A. Assessing support networks: stability and evidence for convergent and divergent validity. *Am J Community Psychol* 1988;**16**:793–809.

Jessmer SL, Anderson D. The effect of politeness and grammar on user perceptions of electronic mail. *N Am J Psychol* 2001;**3**:331–46.

Job RFS, Bullen RB. The effects of a face to face interview versus a group administered questinaire in determinig reaction to noise in the workplace. *J Sound Vib* 1987;**116**:161–8.

Joinson A. Social desirability, anonymity, and Internet-based questionnaires. *Behav Res Methods* 1999;**31**:433–8.

Jones D, Kazis L, Lee A, Rogers W, Skinner K, Cassar L, *et al.* Health status assessments using the Veterans SF-12 and SF-36: methods for evaluating otucomes in the Veterans Health Administration. *JACM* 2001;**24**:68–86.

Jones JW, Brasher EE, Huff JW. Innovations in integrity-based personnel selection: building a technology-friendly assessment. *Int J Select Assess* 2002;**10**:87–97.

Jordon LA, Marcus AC, Reeder LG. Response styles in telephone and household interviewing: a field experiment. *Pub Opin Q* 1980;**44**:210–22.

Jorge MR, Masur J. The use of the short-form Alcohol Dependence Data questionnaire (SADD) in Brazilian alcoholic patients. *Br J Addict* 1985;**80**:301–5.

Kabzems V, Das J. Assessment of extraversion and neuroticism for mentally retarded persons: comparison between questionnaire and video formats. *Dev Disabil Bull* 1990;**18**:20–35.

Kantor J. The effects of computer administration and identification on the Job Descriptive Index (JDI). *J Bus Psychol* 1991;**5**:309–23.

Kapes JT, Vansickle TR. Comparing paper-pencil and computer-based versions of the Harrington-O'Shea Career Decision Making System. *Meas Eval Couns Dev* 1992;**25**:13 May.

Kaplan CP, Hilton JF, Park-Tanjasiri S, Perez-Stable EJ. The effect of data collection mode on smoking attitudes and behavior in young African American and Latina women. Face-to-face interview versus self-administered questionnaires. *Eval Rev* 2001;**25**:454–73.

Kaplan CP, Tanjasiri SP. The effects of interview mode on smoking attitudes and behavior: self-report among female Latino adolescents. *Subst Use Misuse* 1996;**31**:947–63.

Kaplan ML, Asnis GM, Sanderson WC, Keswani L, Lecuona JMD, Joseph S. Suicide assessment: clinical interview vs. self-report. *J Clin Psychol* 1994;**50**:294–8.

Kaplan RM, Sieber WJ, Ganiats TG. The Quality of Well-being Scale: comparison of the interviewer-administered version with a self-administered questionnaire. *Psychol Health* 1997;**12**:783–91.

Kennedy T, Jones R. Development of a postal health status questionnaire to identify people with dyspepsia in the general population. *Scand J Prim Health* 1995;**13**:243–9.

Kiesler S, Sproull LS. Response effects in the electronic survey. *Pub Opin Q* 1986;**50**:402–13.

Kincade KM, Kleine PF, Vaughn J. Methodological issues in the assessment of children's reading interests. *J Instruct Psychol* 1993;**20**:224–36.

King WC, Miles EW. A quasi-experimental assessment of the effect of computerizing noncognitive paper-and-pencil measurements: a test of measurement equivalence. *J Appl Psychol* 1995;**80**:643–51.

Kirsch AD, McCormack MT, Saxon-Harrold SKE. Evaluation of differences in giving and volunteering data collected by in-home and telephone interviewing. *NVSQ* 2001;**30**:495–504.

Klemm P, Hardie T. Depression in internet and face-to-face cancer support groups: a pilot study. *Oncol Nurs Forum* 2002;**29**:E45–51.

Klepac RK, Dowling J, Rokke P, Dodge L, Schafer L. Interview vs. paper-and-pencil administration of the McGill Pain Questionnaire. *Pain* 1981;**11**:241–6.

Klinkenberg W, Calsyn RJ, Morse GA, McCudden S, Richmond T, Burger GK, *et al.* Effect of data collection mode on self-reported sexual and drug using behaviors for persons with severe mental illness. *Eval Program Plann* 2003;**26**:275–82.

Kobak KA, Schaettle SC, Greist JH, Jefferson JW, Katzelnick DJ, Dottl SL. Computer-administered rating scales for social anxiety in a clinical drug trial. see comment. *Depress Anxiety* 1998;**7**:97–104.

Kongerud J, Vale JR, Aalen OO. Questionnaire reliability and validity for aluminum potroom workers. *Scand J Work Environ Health* 1989;**15**:364–70.

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. *Can J Rehabil* 1993;**7**:73–5.

Korner-Bitensky N, Wood-Dauphinee S, Shapiro S, Becker R. Eliciting health status information by telephone after discharge from hospital: Health professionals versus trained lay persons. *Can J Rehabil* 1994;**8**:23–34.

Korner-Bitensky N, Wood-Dauphinee S, Siemiatycki J, Shapiro S, Becker R. Health-related information postdischarge: telephone versus face-to-face interviewing. *Arch Phys Med Rehabil* 1994;**75**:1287–96.

Kraus L, Augustin R. Measuring alcohol consumption and alcohol-related problems: comparison of responses from self-administered questionnaires and telephone interviews. *Addiction* 2001;**96**:459–71.

Krysan M, Schuman H, Scott LJ, Beatty P. Response rates and response content in mail versus face-to-face surveys. *Pub Opin Q* 1994;**58**:381–99.

Kuran T, McCaffery EJ. Expanding discrimination research: beyond ethnicity and to the web. *Soc Sci Q* 2004;**85**:713–30.

Kurt R, Bogner HR, Straton JB, Tien AY, Gallo JJ. Computer-assisted assessment of depression and function in older primary care patients. *Comput Meth Programs Biomed* 2004;**73**:165–71.

Labarere J, Francois P, Bertrand D, Fourny M, Olive F, Peyrin JC. Evaluation of inpatient satisfaction. Comparison of different survey methods. *Presse Med* 2000;**29**:1112–14.

Laird G, Wiebe E, Pulliam P, Thalji L, Huggins V. Assessment of mode-effects in a web-enabled study of civic attitudes. *J Am Stat Assoc* 2002:1983–8.

Lambert ME, Andrews RH, Rylee K, Skinner JR. Equivalence of computerized and traditional MMPI administration with substance abusers. *Comput Hum Behav* 1987;**3**:139–43.

Lankford J, Bell RW, Elias JW. Computerized versus standard personality measures: equivalency, computer anxiety, and gender differences. *Comput Hum Behav* 1994;**10**:497–510.

Lauritsen K, Innocenti AD, Hendel L, Praest J, Lytje MF, Clemmensen-Rotne K, *et al.* Symptom recording in a randomised clinical trial: paper diaries vs. electronic or telephone data capture. *Contr Clin Trials* 2004;**25**:585–97.

Lautenschlager GJ, Flaherty VL. Computer administration of questions: more desirable or more social desirability? *J Appl Psychol* 1990;**75**:310–14.

Leggett CG, Kleckner NS. Social desirability bias in contingent valuation surveys administered through. *Land Econ* 2003;**79**:561–75.

Leidy NK, Elixhauser A, Rentz AM, Beach R, Pellock J, Schachter S, *et al.* Telephone validation of the Quality of Life in Epilepsy Inventory-89 (QOLIE-89). *Epilepsia* 1999;**40**:97–106.

Lenert LA, Hornberger JC. Computer-assisted quality of life assessment for clinical trials. Proceedings/AMIA Annual Fall Symposium 1996:992–6.

Lewis B, Lewis D, Cumming G. Frequent measurement of chronic pain: an electronic diary and empirical findings. *Pain* 1995;**60**:341–7.

Liefeld JP. Response effects in computer-administered questioning. *J Mark Research* 1988;**25**:405–9.

Ljubotina D, Muslic L. Convergent validity of four instruments for measuring posttraumatic stress disorder. *Rev Psychol* 2003;**10**:21.

Llabre MM, Clements NE, Fitzhugh KB, Lancelotta G. The effect of computer-administered testing on test anxiety and performance. *J Educ Comput Res* 1987;**3**:429–33.

Locke SD, Gilbert BO. Method of psychological assessment, self-disclosure, and experiential differences: a study of computer, questionnaire, and interview assessment formats. *J Soc Behav Pers* 1995;**10**:255–63.

Loomis J, King M. Comparison of mail and telephone-mail contingent valuation surveys. *J Environ Manag* 1994;**41**:309–24.

Lorenz F, Ryan VD. Experiments in general/specific questions: comparing results of mail and telephone surveys. *J Am Stat Assoc* 1996:611–13.

Lorig K, Gonzalez VM, Ritter P, Brey VNd. Comparison of three methods of data collection in an urban Spanish-speaking population. *Nursing Res* 1997;**46**:230–4.

Ludemann R, Watson DI, Jamieson GG. Influence of follow-up methodology and completeness on apparent clinical outcome of fundoplication. *Am J Surg* 2003;**186**:143–7.

Lumsden J, Sampson JP, Reardon RC, Lenz JG, Peterson GW. A comparison study of the paper-and-pencil, personal computer, and internet versions of Holland's Self-Directed Search. *Meas Eval Couns Dev* 2004;**37**:85–93.

Lynch S, Curran S, Montgomery S, Fairhurst D, Clarkson P, Suresh R, *et al.* The Brief Depression Scale – reliability and validity of a new self-rating depression scale. *Prim Care Psychiatr* 2000;**6**:111–18.

Lyons RA, Wareham K, Lucas M, Price D, Williams J, Hutchings HA. SF-36 scores vary by method of administration: implications for study design. *J Publ Health Med* 1999;**21**:41–5.

Maitland ME, Mandel AR. A client-computer interface for questionnaire data. *Arch Phys Med Rehabil* 1994;**75**:639–42.

Margo A, Johnson C, Ancill R, Carr T. Assessment of depression by microcomputer. *Acta Psychiatr Scand* 1983;**67**:434–5.

Marsden J, Jones RB. Validation of web-based questionnaires regarding osteoporosis prevention in young British women. *Health Bull* 2001;**59**:254–62.

Martin CL, Nagao DH. Some effects of computerized interviewing on job applicant responses. *J Appl Psychol* 1989;**74**:72–80.

McColl MA, Paterson M, Davies D, Doubt L, Law M. Validity and community utility of the Canadian Occupational Performance Measure. *Can J Occup Ther* 2000;**67**:22–30.

McDevitt PK, Small MH. Proprietary market research: are online panels appropriate? Market Intell Plan 2002;**20**:285–96.

McDonagh EC, Rosenblum AL. A comparison of mailed questionnaires and subsequent structured interviews. *Pub Opin Q* 1965;**29**:131–6.

McFarlane J, Christoffel K, Bateman L, Miller V, Bullock L. Assessing for abuse: self-report versus nurse interview. *Publ Health Nurse* 1991;**8**:245–50.

Menon A, Kondapavalru P, Krishna P, Chrismer J, Raskin A, Hebel J, *et al.* Evaluation of a portable low cost videophone system in the assessment of depressive symptoms and cognitive function in elderly medically ill veterans. *J Nerv Ment Dis* 2001;**189**:399–401.

Merten T, Ruch W. A comparison of computerized and conventional administration of the German versions of the Eysenck Personality Questionnaire and the Carroll Rating Scale for Depression. *Pers Indiv Differ* 1996;**20**:281–91.

Merten T, Siebert K. A comparison of computerized and conventional administration of the EPQ-R and CRS: further data on the Merten and Ruch (1996) study. *Pers Indiv Differ* 1997;**22**:283–6.

Metsahonkala L, Sillanpaa M, Tuominen J. Headache diary in the diagnosis of childhood migraine. *Headache* 1997;**37**:240–4.

Metzger DS, Koblin B, Turner C, Navaline H, Valenti F, Holte S, *et al.* Randomized controlled trial of audio computer-assisted self-interviewing: utility and acceptability in longitudinal studies. HIVNET Vaccine Preparedness Study Protocol Team. *Am J Epidemiol* 2000;**152**:99–106.

Meyer N, Fischer R, Weitkunat R, Crispin A, Schotten K, Bellach BM, *et al.* Evalutation of health monitoring in Bavaria by computer-assisted telephone interviews (CATI) in comparison to the German National Health Examination Survey conducted in 1998 by the Robert Koch Institute. *Gesundheitswesen* 2002;**64**:329–36.

Midanik LT, Greenfield TK. Telephone versus in-person interviews for alcohol use: results of the 2000 National Alcohol Survey. *Drug Alcohol Depend* 2003;**72**:209–14.

Midanik LT, Greenfield TK, Rogers JD. Reports of alcohol-related harm: telephone versus face-to-face interviews. *J Stud Alcohol* 2001;**62**:74–8.

Miller ET, Neal DJ, Roberts LJ, Baer JS, Cressler SO, Metrik J, *et al.* Test-retest reliability of alcohol measures: is there a difference between internet-based assessment and traditional methods? Psychol Addict Behav2002;**16**:56–63.

Miller TI, Kobayashi MM, Caldwell E, Thurston S, Collett B. Citizen surveys on the web: general population surveys of community opinion. *Soc Sci Comput Rev* 2002;**20**:124–36.

Mills JF, Kroner DG, Forth AE. Novaco Anger Scale: reliability and validity within an adult criminal sample. *Assessment* 1998;**5**:237–48.

Millstein SG. Acceptability and reliability of sensitive information collected via computer interview. *Educ Psychol Meas* 1987;**47**:523–33.

Minnick A, Young WB. Comparison between reports of care obtained by postdischarge telephone interview and predischarge personal interview. *Outcome Manag Nurs Pract* 1999;**3**:32–7.

Molitor F, Kravitz RL, To Y, Fink A. Methods in survey research: evidence for the reliability of group administration vs personal interviews. *Am J Public Health* 2001;**91**:826–7.

Moore DS, Cook-Hubbard K. Comparison of methods for evaluating patient response to nursing care. *Nursing Res* 1975;**24**:202–4.

Morishita L, Boult C, Ebbitt B, Rambel M, Fallstrom K, Gooden T. Concurrent validity of administering the Geriatric Depression Scale and the physical functioning dimension of the SIP by telephone. *J Am Geriatr Soc* 1995;**43**:680–3.

Mosley-Williams A, Williams CA. Validation of a computer version of the American College of Rheumatology patient assessment questionnaire for the autonomous self-entry of self-report data in an urban rheumatology clinic. *Arthritis Rheum* 2004;**50**:332–3.

Mulkey LM, Anderson TD. A computer application for research on gender: using online context as a mediating variable in the investigation of sex-role orientation and care-oriented moral reasoning. *Soc Sci Comput Rev* 2002;**20**:137–48.

Mundt JC, Bohn MJ, King M, Hartley MT. Automating standard alcohol use assessment instruments via interactive voice response technology. *ACER* 2002;**26**:207–11.

Munoz RF, McQuaid JR, Gonzalez GM, Dimas J, Rosales VA. Depression screening in a women's clinic: using automated Spanish- and English-language voice recognition. *J Consult Clin Psychol* 1999;**67**:502–10.

Murrelle L, Bulger JD, Ainsworth BE, Holliman SC, Bulger DW. Computerized mental health risk appraisal for college students: user acceptability and correlation with standard pencil-and-paper questionnaires. *Am J Health Promot* 1992;**7**:90–2.

Nair B, Ying X, Maetzel A, Li L, Pencharz J, Maguire L, *et al.* A randomized comparison of telephone and mailed health status questionaire in patients with rheumatoid arthritis. *Arthritis Rheum* 2002;**46**:S114.

Nass C, Moon Y, Carney P. Are people polite to computers? Responses to computer-based interviewing systems. *J Appl Soc Psychol* 1999;**29**:1093–110.

Newman JC, Jarlais DCD, Turner CF, Gribble J, Cooley P, Paone D. The differential effects of face-to-face and computer interview modes. *Am J Public Health* 2002;**92**:294–7.

Nyholm D, Kowalski J, Aquilonius SM. Wireless real-time electronic data capture for self-assessment of motor function and quality of life in Parkinson's disease. *Mov Disord* 2004;**19**:446–51.

O'Dell WF. Personal interviews or mail panels? J Mark 1962;**26**:34–9.

Oei TI, Zwart FM. The assessment of life events: self-administered questionnaire versus interview. *J Affect Disord* 1986;**10**:185–90.

Okamoto K, Ohsuka K, Shiraishi T, Hukazawa E, Wakasugi S, Furuta K. Comparability of epidemiological information between self- and interviewer-administered questionnaires. *J Clin Epidemiol* 2002;**55**:505–11.

Okazaki S. Asian American and White American differences on affective distress symptoms: do symptom reports differ across reporting methods? *J Cross Cult Psychol* 2000;**31**:603–25.

O'Neill D, Rice I, Blake P, Walsh J. The Geriatric Depression Scale: rater-administered or self-administered? Int J Geriatr Psychiatr 1992;**7**:511–15.

Ooijen Mv, Ivens UI, Johansen C, Skov T. Comparison of a self-administered questionnaire and a telephone interview of 146 Danish waste collectors. *Am J Ind Med* 1997;**31**:653–8.

Palmer R, Keyser D. Automated psychological testing with psychiatric patients. *Int J Partial Hosp* 1984;**2**:275–81.

Parks BT, Mead D, Johnson BL. Validation of a computer administered Marital Adjustment Test. *J Martial Fam Ther* 1985;**11**:207–10.

Pasveer KA, Ellard JH. The making of a personality inventory: help from the WWW. *Behav Res Meth Instrum Comput* 1998;**30**:309–13.

Pelissolo A, Lepine J. Validation study of the French version of the TCI. *Ann Med Psychol* 1997;**155**:497–508.

Pelissolo A, Veysseyre O, Lepine JP. Validation of a computerized version of the temperament and character inventory (TCI) in psychiatric inpatients. *Psychiatr Res* 1997;**72**:195–9.

Pendleton D, Wakeford R. Studying medical opinion: a comparison of telephone interviews and postal questionnaires to general practitioners. *Community Med* 1987;**9**:25–34.

Penny JA. Exploring differential item functioning in a 360-degree assessment: rater source and method of delivery. *ORM* 2003;**6**:61–9.

Perdrizet S, Amphoux M, Liard R, Ballereau M, Besch N, Ballu M, *et al.* Respiratory pathology in occupational medicine: evaluation of 3 methods of data collection and research on risk factors. *Rev Mal Respir* 1984;**1**:99–103.

Perkins G, Yuan H. A Comparison of web-based and paper-and-pencil library satisfaction. *Coll Res Lir* 2001;**62**:369–77.

Perkins JJ, Sanson-Fisher RW. An examination of self- and telephone-administered modes of administration for the Australian SF-36. *J Clin Epidemiol* 1998;**51**:969–73.

Peterson L, Johannsson V, Carlsson SG. Computerized testing in a hospital setting: psychometric and psychological effects. *Comput Hum Behav* 1996;**12**:339–50.

Pettit FA. A comparison of World Wide Web and paper-and-pencil personality questionnaires. *Behav Res Methods* 2002;**34**:50–4.

Picavet HS, Van Den Bos GA. Comparing survey data on functional disability: the impact of some methodological differences. *J Epidemiol Community Health* 1996;**50**:86–93.

Pinsoneault TB. Equivalency of computer-assisted and paper-and-pencil administered versions of the Minnesota Multiphasic Personality Inventory-2. *Comput Hum Behav* 1996;**12**:291–300.

Plante M, Corcos J, Gregoire I, Belanger MF, Brock G, Rossingol M. The international prostate symptom score: physician versus self-administration in the quantification of symptomatology. *Urology* 1996;**47**:326–8.

Ployhart RE, Weekley JA, Holtz BC, Kemp C. Web-based and paper-and-pencil testing of applicants in a proctored setting: are personality, biodata and situational judgment tests comparable? Person Psychol 2003;**56**:733–52.

Pomerleau CS, Carton SM, Lutzke ML, Flessland KA, Pomerleau OF. Reliability of the Fagerstrom Tolerance Questionnaire and the Fagerstrom Test for nicotine dependence. *Addict Behav* 1994;**19**:33–9.

Pouwer F, Snoek FJ, Ploeg HMvd, Heine RJ, Brand AN. A comparison of the standard and the computerized versions of the Well-being Questionnaire (WBQ) and the Diabetes Treatment Satisfaction Questionnaire (DTSQ). *Qual Life Res* 1998;**7**:33–8.

Pruchno RA, Hayden JM. Interview modality: effects on costs and data quality in a sample of older women. *J Ageing Health* 2000;**12**:24 March.

Pugh N, Iannacchione V, Lance T, Dimitropoulos L. Evaluating Mode Effects in the Medicare CAHPS® Fee-for-Service Survey. *J Am Stat Assoc* 2002:272–6, section on survey research.

Puhan MA, Behnke M, Laschke M, Lichtenschopf A, Brandli O, Guyatt GH, *et al.* Self-administration and standardisation of the chronic respiratory questionnaire: a randomised trial in three German-speaking countries. *Respir Med* 2004;**98**:342–50.

Pyne JM, Sieber WJ, David K, Kaplan RM, Rapaport MH, Williams DK. Use of the quality of well-being self-administered version (QWB-SA) in assessing health-related quality of life in depressed patients. *J Affect Disord* 2003;**76**:237–47.

Quine S. 'Does the mode matter?': a comparison of three modes of questionnaire completion. *Community Health Stud* 1985;**9**:151–6.

Quinn P, Goka J, Richardson H. Assessment of an electronic daily diary in patients with overactive bladder. *BJU Int* 2003;**91**:647–52.

Rabin JM, McNett J, Badlani GH. Compu-Void II': the computerized voiding diary. *J Med Syst* 1996;**20**:19–34.

Radloff LS. The CES-D Scale: s self-report depression scale for research in the general population. *Appl Psychol Meas* 1977;**1**:385–401.

Rakowski W, Julius M, Hickey T, Verbrugge LM, Halter JB. Daily symptoms and behavioral responses. Results of a health diary with older adults. *Med Care* 1988;**26**:278–97.

Reilly WT, Talley NJ, Pemberton JH, Zinsmeister AR. Validation of a questionnaire to assess fecal incontinence and associated risk factors: Fecal Incontinence Questionnaire. *Dis Colon Rectum* 2000;**43**:146–53; discussion 153–4.

Revicki DA, Tohen M, Gyulai L, Thompson C, Pike S, Davis-Vogel A, *et al.* Telephone versus in-person clinical and health status assessment interviews in patients with bipolar disorder. *Harv Rev Psychiatr* 1997;**5**:75–81.

Rhee KJ, Allen RA, Bird J. Telephone vs mail response to an emergency department patient satisfaction survey. *Acad Emerg Med* 1998;**5**:1121–3.

Rhodes T, Girman CJ, Jacobsen SJ, Guess HA, Hanson KA, Oesterling JE, *et al.* Does the mode of questionnaire administration affect the reporting of urinary symptoms? Urology 1995;**46**:341–5.

Ridgway J, MacCulloch M, Mills H. Some experiences in administering a psychometric test with a light pen and microcomputer. *Int J Man Mach Stud* 1982;**17**:265–78.

Riva G, Teruzzi T, Anolli L. The use of the internet in psychological research: comparison of online and offline questionnaires. *Cyberpsychol Behav* 2003;**6**:73–80.

Roberts LL, Konczak LJ, Macan TH. Effects of data collection method on organizational climate survey results. *Appl HRM Res* 2004;**9**:13–26.

Rombouts R, Gazendam A, Nijholt M. Study of the equivalence of computer-assisted vs paper-and-pencil versions of some psychological questionnaires. *Ned Tijdschr Psychol* 1989;**44**:88–93.

Romer D, Hornik R, Stanton B, Black M, Li X, Ricardo I, *et al.* Talking computers: a reliable and private method to conduct interviews on sensitive topics with children. *J Sex Res* 1997;**34**:3 September.

Rose M, Hess V, Horhold M, Brahler E, Klapp BF. Mobile computer-assisted psychometric diagnosis. Economic advantages and results on test stability. *Psychother Psychosom Med Psychol* 1999;**49**:202–7.

Rosenfeld P, Booth-Kewley S, Edwards JE, Thomas MD. Responses on computer surveys: impression management, social desirability, and the Big Brother syndrome. *Comput Hum Behav* 1996;**12**:263–74.

Rosenfeld P, Giacalone RA, Knouse SB, Doherty LM. Impression management, candor, and microcomputer-based organizational surveys: an individual differences approach. *Comput Hum Behav* 1991;**7**:23–32.

Ross MW, Tikkanen R, Mansson SA. Differences between internet samples and conventional samples of men who have sex with men: implications for research and HIV interventions. *Soc Sci Med* 2000;**51**:749–58.

Rossier P, Wade DT. The Guy's Neurological Disability Scale in patients with multiple sclerosis: a clinical evaluation of its reliability and validity. *Clin Rehabil* 2002;**16**:75–95.

Roster CA, Rogers RD, Albaum G, Klein D. A comparison of response characteristics from web and telephone surveys. *Int J Market Res* 2004;**46**:359–73.

Rozensky RH, Honor LF, Rasinski K, Tovian SM. Paper-and-pencil versus computer-administered MMPIs: a comparison of patients' attitudes. *Comput Hum Behav* 1986;**2**:111–16.

Rubenach S, Anderson CS, Laubscher S. The Short Form-12 by telephone as a measure of health-related quality of life after stroke. *Age Ageing* 2000;**29**:553–4.

Russell GG, Flight I, Leppard P, Pabst JAvLv. A comparison of paper-and-pencil and computerised methods of 'hard' laddering. *Food Qual Prefer* 2004;**15**:279–91.

Ryan JM, Corry JR, Attewell R, Smithson MJ. A comparison of an electronic version of the SF-36 General Health Questionnaire to the standard paper version. *Qual Life Res* 2002;**11**:19–26.

Saameno JAB, Sanchez AD, Castillo JDLd, Claret PL. Validity and reliability of the Duke-UNC-11 questionnaire of functional social support. *Aten Prim* 1996;**18**:153–6.

Saameno JAB, Sanchez AD, Castillo JDLd, Claret PL. Validity and reliability of the family Apgar family function test. *Aten Prim* 1996;**18**:289–96.

Saleh KJ, Radosevich DM, Kassim RA, Moussa M, Dykes D, Bottolfson H, *et al.* Comparison of commonly used orthopaedic outcome measures using palm-top computers and paper surveys. *J Orthop Surg Res* 2002;**20**:1146–51.

Salgado JF, Moscoso S. Internet-based personality testing: equivalence of measures and assessees' perceptions and reactions. *Int J Select Assess* 2003;**11**:194–205.

Sanitioso R, Reynolds JH. Comparability of standard and computerized administration of two personality questionnaires. *Pers Indiv Differ* 1992;**13**:899–907.

Scandell DJ, Wlazelek B. Validation study of the AIDS Health Belief Scale. *Can J Hum Sex* 2002;**11**:41–9.

Schmitz N, Hartkamp N, Brinschwitz C, Michalek S. Computerized administration of the Symptom Checklist (SCL-90-R) and the Inventory of Interpersonal Problems (IIP-C) in psychosomatic outpatients. *Psychiatr Res* 1999;**87**:217–21.

Schmitz N, Hartkamp N, Brinschwitz C, Michalek S, Tress W. Comparison of the standard and the computerized versions of the Symptom Check List (SCL-90-R): a randomized trial. *Acta Psychiatr Scand* 2000;**102**:147–52.

Schonlau M, Zapert K, Simon LP, Sanstad KH, Marcus SM, Adams J, *et al.* A comparison between responses from a propensity-weighted web survey and an identical RDD survey. *Soc Sci Comput Rev* 2004;**22**:128–38.

Schuldberg D. The MMPI is less sensitive to the automated testing format than it is to repeated testing: item and scale effects. *Comput Hum Behav* 1988;**4**:285–98.

Schulenberg SE, Yutrzenka BA. Equivalence of computerized and conventional versions of the Beck Depression Inventory-II (BDI-II). *Curr Psychol* 2001;**20**:216–30.

Schwartz SJ, Mullis RL, Dunham RM. Effects of authoritative structure in the measurement of identity formation: individual computer-managed versus group paper-and-pencil testing. *Comput Hum Behav* 1998;**14**:239–48.

Schwarz N, Hippler HJ. The numeric values of rating scales: a comparison of their impact in mail surveys and telephone interviews. *Int J Publ Opin Res* 1995;**7**:72–4.

Scissons EH. Computer administration of the California Psychological Inventory. *Meas Eval Guid* 1976;**91**:22–5.

Shannon DM, Bradshaw CC. A comparison of response rate, response time, and costs of mail and electronic surveys. *J Exp Educ* 2002;**70**:179–92.

Sieber WJ, David KM, Adams JE, Kaplan RM, Ganiats TG. Assessing the impact of migraine on health-related quality of life: an additional use of the quality of well-being scale-self-administered. *Headache* 2000;**40**:662–71.

Siemiatycki J. A comparison of mail, telephone, and home interview strategies for household health surveys. *Am J Public Health* 1979;**69**:238–45.

Simola SK, Holden RR. Equivalence of computerized and standard administration of the Piers-Harris Children's Self-Concept Scale. *J Pers Assess* 1992;**58**:287–94.

Simon GE, Revicki D, VonKorff M. Telephone assessment of depression severity. *J Psychiatr Res* 1993;**27**:247–52.

Skinner HA, Allen BA. Does the computer make a difference? Computerized versus face-to-face versus self-report assessment of alcohol, drug, and tobacco use. *J Consult Clin Psychol* 1983;**51**:267–75.

Smeeth L, Fletcher AE, Stirling S, Nunes M, Breeze E, Ng E, *et al.* Randomised comparison of three methods of administering a screening questionnaire to elderly people: findings from the MRC trial of the assessment and management of older people in the community. *BMJ* 2001;**323**:1403–7.

Smither JW, Walker AG, Yap MK. An examination of the equivalence of web-based versus paper-and-pencil upward feedback ratings: rater- and ratee-level analyses. *Educ Psychol Meas* 2004;**64**:40–61.

Snyder-Ramos SA, Seintsch H, Bottiger BW, Motsch J, Martin E, Bauer M. Patient satisfaction and information gain after the preanesthetic visit: a comparison of face-to-face interview, brochure, and video'. *Anesthesia and analgesia* 2005;**100**:1753–8.

Sorensen S, Rylander R, Berglund K. Interviews and mailed questionnaires for the evaluation of annoyance reactions. *Environ Res* 1974;**8**:166–70.

Sparrow N, Curtice J. Measuring the attitudes of the general public via Internet polls: an evaluation. *Int J Market Res* 2004;**46**:23–44.

Speer DC. An evaluation of the Denver Community Mental Health Questionnaire as a measure of outpatient treatment effectiveness. *Eval Q* 1977;**1**:475–92.

Stanton JM. An empirical assessment of data collection using the internet. *Person Psychol* 1998;**51**:709–25.

Stapleton CN, Norris S, Brady S. Customer satisfaction with internet and IVR as census data collection tools. *J Am Stat Assoc* 2003:4039–46.

Stones MJ, Kozma A. Multidimensional assessment of the elderly via a microcomputer: the SENOTS program and battery. *Psychol Ageing* 1989;**4**:113–18.

St-Pierre M, Beland Y. Mode effects in the Canadian Community Health Survey: a comparison of CAPI and CATI. *J Am Stat Assoc* 2005:4438–45.

Stratton RJ, Stubbs RJ, Hughes D, King N, Blundell JE, Elia M. Comparison of the traditional paper visual analogue scale questionnaire with an Apple Newton electronic appetite rating system (EARS) in free living subjects feeding ad libitum. *Eur J Clin Nutr* 1998;**52**:737–41.

Strayer M, Kuthy R, Sutton S. Elderly nonrespondents to a mail survey: a telephone follow-up. *Spec Car Dent* 1993;**13**:245–8.

Stringfellow VL, Fowler FJ, Clarridge BR. Evaluating mode effects on a survey of behavioral health care users. *J Am Stat Assoc* 2001:1 May.

Stubbs RJ, Hughes DA, Johnstone AM, Rowley E, Ferris S, Elia M, *et al.* Description and evaluation of a Newton-based electronic appetite rating system for temporal tracking of appetite in human subjects. *Physiol Behav* 2001;**72**:615–19.

Supple AJ, Aquilino WS, Wright DL. Collecting sensitive self-report data with laptop computers: impact on the response tendencies of adolescents in a home interview. *J Res Adolesc* 1999;**9**:467–88.

Swanston M, Abraham C, Macrae WA, Walker A, Rushmer R, Elder L, *et al.* Pain assessment with interactive computer animation. *Pain* 1993;**53**:347–51.

Synodinos NE, Papacostas C, Okimoto GM. Computer-administered versus paper-and-pencil surveys and the effect of sample selection. *Behav Res Meth Instrum Comput* 1994;**26**:395–401.

Taenzer P, Bultz BD, Carlson LE, Speca M, DeGagne T, Olson K, *et al.* Impact of computerized quality of life screening on physician behaviour and patient satisfaction in lung cancer outpatients. Psycho-Oncology 2000;**9**:203–13.

Taenzer PA, Speca M, Atkinson MJ, Bultz BD, Page S, Harasym P, *et al.* Computerized quality-of-life screening in an oncology clinic. *Canc Pract* 1997;**5**:168–75.

Talley JE, Barrow JC, Fulkerson KF, Moore CA. Conducting a needs assessment of university psychological services: a campaign of telephone and mail strategies. *J Am Coll Health* 1983;**32**:101–3.

Talley NJ, Boyce PM, Owen BK, Newman P, Paterson KJ. Initial validation of a bowel symptom questionnaire and measurement of chronic gastrointestinal symptoms in Australians. *Aust New Zeal J Med* 1995;**25**:302–8.

Taylor H. Does internet research work. *Int J Market Res* 2000;**42**:51–62.

Theiler R, Bischoff-Ferrari HA, Good M, Bellamy N. Responsiveness of the electronic touch screen WOMAC 3.1 OA Index in a short term clinical trial with rofecoxib. *Osteoarthritis Cartilage* 2004;**12**:912–16.

Theiler R, Spielberger J, Bischoff HA, Bellamy N, Huber J, Kroesen S. Clinical evaluation of the WOMAC 3.0 OA Index in numeric rating scale format using a computerized touch screen version. *Osteoarthritis Cartilage* 2002;**10**:479–81.

Thomsen JF, Mikkelsen S. Interview data versus questionnaire data in the diagnosis of carpal tunnel syndrome in epidemiological studies. *Occup Med* 2003;**53**:57–63.

Tittler BI, Weitz LJ, Anchor KN. Pretest and change score intercorrelations in the validation of behavioral measures of openness. *J Clin Psychol* 1976;**32**:806–8.

Truman J, Robinson K, Evans AL, Smith D, Cunningham L, Millward R, *et al.* The Strengths and Difficulties Questionnaire: a pilot study of a new computer version of the self-report scale. *Eur Child Adolesc Psychiatry* 2003;**12**:14 September.

Tryon WW, Orr DA, Blumenfield M. Psychometric equivalence of an electronic visual-analog(EVA): a conventinal visual-analog and a likert rating scale. *Int J Meth Psychiatr Res* 1996;**6**:123–7.

Tseng HM, Tiplady B, Macleod HA, Wright P. Computer anxiety: a comparison of pen-based personal digital assistants, conventional computer and paper assessment of mood and performance. *Br J Psychol* 1998;**89**:599–610.

Unruh M, Yan G, Radeva M, Hays RD, Benz R, Athienites NV, *et al.* Bias in assessment of health-related quality of life in a hemodialysis population: a comparison of self-administered and interviewer-administered surveys in the HEMO study. *J Am Soc Nephrol* 2003;**14**:2132–41.

Van Den Kerkhof EG, Parlow JL, Goldstein DH, Milne B. In Canada, anesthesiologists are less likely to respond to an electronic, compared to a paper questionnaire. *Can J Anesth* 2004;**51**:449–54.

Vansickle TR, Kapes JT. Comparing paper-pencil and computer-based versions of the Strong-Campbell Interest Inventory. *Comput Hum Behav* 1993;**9**:441–9.

Vaske D, Whittaker JJ. Mail versus telephone surveys: potential biases in expenditure and willingness-to-pay data. *JPRA* 1998;**16**:15–30.

Velde At, Sprangers MA, Aaronson NK. Feasibility, psychometric performance, and stability across modes of administration of the CARES-SF. *Ann Oncol* 1996;**7**:381–90.

Velikova G, Wright EP, Smith AB, Cull A, Gould A, Forman D, *et al.* Automated collection of quality-of-life data: a comparison of paper and computer touch-screen questionnaires. *J Clin Oncol* 1999;**17**:998–1007.

Vereecken C. Paper pencil versus pc administered querying of a study on health behaviour in school-aged children. *Arch Publ Health* 2001;**59**:43–61.

Verrips GH, Vogels AG, Ouden ALd, Paneth N, Verloove-Vanhorick SP. Measuring health-related quality of life in adolescents: agreement between raters and between methods of administration. *Child Care Health Dev* 2000;**26**:457–69.

Vispoel WP, Boo J, Bleiler T. Computerized and paper-and-pencil versions of the Rosenberg Self-Esteem Scale: a comparison of psychometric features and respondent preferences. *Educ Psychol Meas* 2001;**61**:461–74.

Walker AH, Restuccia JD. Obtaining information on patient satisfaction with hospital care: mail versus telephone. *Health Serv Res* 1984;**19**:291–306.

Warner JL, Berman JJ, Weyant JM, Ciarlo JA. Assessing mental health program effectiveness: a comparison of three client follow-up methods. *Eval Rev* 1983;**7**:635–58.

Webster J, Compeau D. Computer-assisted versus paper-and-pencil administration of questionnaires. *Behav Res Meth Instrum Comput* 1996;**28**:567–76.

Weinberger M, Nagle B, Hanlon JT, Samsa GP, Schmader K, Landsman PB, *et al.* Assessing health-related quality of life in elderly outpatients: telephone versus face-to-face administration. *J Am Geriatr Soc* 1994;**42**:1295–9.

Weinberger M, Oddone EZ, Samsa GP, Landsman PB. Are health-related quality-of-life measures affected by the mode of administration? *J Clin Epidemiol* 1996;**49**:135–40.

Weir P, Laurence M, Blessing C. A comparison of the use of telephone interview to telephone audio CASI in a customer satisfaction survey. *J Am Stat Assoc* 2000:828–33.

White DM, Clements CB, Fowler RD. A comparison of computer administration with standard administration of the MMPI. *Comput Hum Behav* 1985;**1**:153–62.

Whitener EM, Klein HJ. Equivalence of computerized and traditional research methods: the roles of scanning, social environment, and social desirability. *Comput Hum Behav* 1995;**11**:65–75.

Whittier DK, Seeley S, Lawrence JSS. A comparison of web- with paper-based surveys of gay and bisexual men who vacationed in a gay resort community. *AIDS Educ Prev* 2004;**16**:476–85.

Wijck EEv, Bosch JL, Hunink MG. Time-tradeoff values and standard-gamble utilities assessed during telephone interviews versus face-to-face interviews. *Med Decis Making* 1998;**18**:400–5.

Wilkerson JM, Nagao DH, Martin CL. Socially desirable responding in computerized questionnaires: when questionnaire purpose matters more than the mode. *J Appl Soc Psychol* 2002;**32**:544–59.

Williams JE, Singh SJ, Sewell L, Guyatt GH, Morgan MD. Development of a self-reported Chronic Respiratory Questionnaire (CRQ-SR). *Thorax* 2001;**56**:954–9.

Williams JE, Singh SJ, Sewell L, Morgan MD. Health status measurement: sensitivity of the self-reported Chronic Respiratory Questionnaire (CRQ-SR) in pulmonary rehabilitation. *Thorax* 2003;**58**:515–18.

Wilson AS, Kitas GD, Carruthers DM, Reay C, Skan J, Harris S, *et al.* Computerized information-gathering in specialist rheumatology clinics: an initial evaluation of an electronic version of the Short Form 36. *Rheumatology* 2002;**41**:268–73.

Wilson F, Genco KT, Yager GG. Assessing the equivalence of paper-and-pencil vs. computerized tests: demonstration of a promising methodology. *Comput Hum Behav* 1985;**1**:265–75.

Wiseman F. Methodological bias in public opinion surveys. *Pub Opin Q* 1972;**36**:105–8.

Wit Rd, Dam Fv, Hanneman M, Zandbelt L, Buuren Av, Heijden Kvd, *et al.* Evaluation of the use of a pain diary in chronic cancer pain patients at home. *Pain* 1999;**79**:89–99.

Woehr DJ, Miller MJ, Lane JAS. The development and evaluation of a computer-administered measure of cognitive complexity. *Pers Indiv Differ* 1998;**25**:1037–49.

Wolffsohn JS, Cochrane AL, Watt NA. Implementation methods for vision related quality of life questionnaires. *Br J Ophthalmol* 2000;**84**:1035–40.

Wright DL, Aquilino WS, Supple AJ. A comparison of computer-assisted paper-and-pencil self-administered questionnaires in a survey on smoking, alcohol, and drug use. *Pub Opin Q* 1998;**62**:331–53.

Wright JG, Young NL, Waddell JP. The reliability and validity of the self-reported patient-specific index for total hip arthroplasty. *J Bone Joint Surg* 2000;**82**:829–37.

Wright L, May K, Jackson K. Exaggerated social control and its relationship to the Type A behavior pattern as measured by the structured interview. *J Res Pers* 1991;**25**:135–6.

Wu AW, Jacobson DL, Berzon RA, Revicki DA, Horst Cvd, Fichtenbaum CJ, *et al.* The effect of mode of administration on medical outcomes study health ratings and EuroQol scores in AIDS. *Qual Life Res* 1997;**6**:3 October.

Yarnold PR, Stewart MJ, Stille FC, Martin GJ. Assessing functional status of elderly adults via microcomputer. *Percept Mot Skills* 1996;**82**:689–90.

Yun GW, Trumbo CW. Comparative response to a survey executed by post, e-mail, & web form. *J Comput Mediat Commun* 2000;**6**(1).

Appendix 6

Papers excluded at second stage

Adams CD, Perkins KC, Lumley V, Hughes C, Burns JJ, Omar HA. Validation of the Perkins Adolescent Risk Screen (PARS). *J Adolesc Health* 2003;**33**:462–70.

Albinsson G. [Custom-made reducing with tape measure and computer. Interview by Viveka Holmertz.] *Vardfacket* 1988;**12**:4–5.

Albons B. [Which interviewers are best: computers or doctors?] *Sjukskoterskan* 1987:25–9.

Alderfer CP. Convergent and discriminant validation of satisfaction and desire measures by interviews and questionnaires. *J Appl Psychol* 1967;**51**:509–20.

Anderson JR, Waldron I. Behavioural and content components of the structured interview assessment of the type A behaviour pattern in women. *J Behav Med* 1983;**6**:123–34.

Aneshensel CS, Yokopenic PA. Tests for the comparability of a causal model of depression under two conditions of interviewing. *J Pers Soc Psychol* 1985;**49**:1337–48.

Angle HV, Johnsen T, Grebenkemper NS, Ellinwood EH. Computer interview support for clinicians. *Prof Psychol Res Pract* 1979;**10**:49–57.

Araki S, Murata K, Yokoyama K, Kawakami N. [Subclinical neuro-psychobehavioral effects in occupational, environmental and community health: methodology and recent findings.] *Nippon Eiseigaku Zasshi* 1995;**50**:713–29.

Arestova O. Computerisation of experiments and validity of psychodiagnostic methods. *Sov J Psychol* 1990;**11**:68–77.

Aronov DM, Kovaleva OF, Aleshin OI, Mazaev VP, Rozhnov AV, Danielov GE. [Possibilities of the questionnaire method for detection of the symptoms of preclinical stages of coronary arteriosclerosis and ischemic heart disease.] *Kardiologiia* 1991;**31**:32–5.

Baker EL, Letz RE, Fidler AT, Shalat S, Plantamura D, Lyndon M. A computer-based neurobehavioral evaluation system for occupational and environmental epidemiology: methodology and validation studies. *Neurobehav Toxicol Teratol* 1985;**7**:369–77.

Bakke PS, Hanoa R, Gulsvik A. Relation of occupational exposure to respiratory symptoms and asthma in a general population sample: self-reported versus interview-based exposure data. *Am J Epidemiol* 2001;**154**:477–83.

Ball CJ, Scott N, McLaren PM, Watson JP. Preliminary evaluation of a Low-Cost VideoConferencing (LCVC) system for remote cognitive testing of adult psychiatric patients. *Br J Clin Psychol* 1993;**32**:303–7.

Beck AT, Steer RA, Ranieri WF. Scale for Suicide Ideation: psychometric properties of a self-report version. *J Clin Psychol* 1988;**44**:499–505.

Beck F, Pertetti-Watel P. Les usages de drogues illicites déclarés par les adolescents selon le mode de collecte. *Population* 2001;**56**:963–86.

Bedard M, Molloy D, Guyatt GH, Standish T. Self-administered and interviewer-administered instruments for dementia research. *Clin Gerontol* 1998;**19**:25–35.

Beebe TJ, Mika T, Harrison PA, Anderson RE. Computerised school surveys: design and development issues. *Soc Sci Comput Rev* 1997;**15**:159–69.

Begg A, Drummond G, Tiplady B. Assessment of postsurgical recovery after discharge using a pen computer diary. *Anaesthesia* 2003;**58**:1101–5.

Berg M. Evaluation of a questionnaire used in dermatological epidemiology. Discrepancy between self-reported symptoms and objective signs. *Acta Dermatol Venereol* 1991;**156**:S13–7.

Berge E, Fjaertoft H, Indredavik B, Sandset PM. Validity and reliability of simple questions in assessing short- and long-term outcome in Norwegian stroke patients. *Cerebrovasc Dis* 2001;**11**:305–10.

Berger PK, Sullivan JE. Instructional set, interview context, and the incidence of don't know responses. *J Appl Psychol* 1970;**54**:414–16.

Bermack E. Effect of telephone and face-to-face communication on rated extent of self-disclosure by female college students. *Psychol Rep* 1989;**65**:259–67.

Bernstein DP, Fink L, Handelsman L, Foote J, Lovejoy M, Wenzel K, *et al.* Initial reliability and validity of a new retrospective measure of child abuse and neglect. *Am J Psychiatr* 1994;**151**:1132–6.

Bethell C, Fiorillo J, Lansky D, Hendryx M, Knickman J. Online consumer surveys as a methodology for assessing the quality of the United States health care system. *J Med Internet Res* 2004;**6**:e2.

Bifulco A, Mahon J, Kwon JH, Moran PM, Jacobs C. The Vulnerable Attachment Style Questionnaire (VASQ): an interview-based measure of attachment styles that predict depressive disorder. *Psychol Med* 2003;**33**:1099–110.

Bill JM. Concurrent and predictive validity of two methods of information gathering: structured interview and questionnaire compared. *Paper read to Annual Meeting of BPS (Northern Ireland Branch)* 1973:88–9.

Blumenthal JA, Herman S, Toole LCO, Haney TL, Williams RB, Jr, Barefoot JC. Development of a brief self-report measure of the type A (coronary prone) behaviour pattern. *J Psychosom Res* 1985;**29**:265–74.

Bolten W, Emmerich M, Weber E, Fassmeyer N. [Validation of electronic by conventional pain diaries.] *Z Rheumatol* 1991;**50**(Suppl. 1):55–64.

Boulger JG. Comparison of two methods of obtaining life history data: structured interview versus questionnaire. *Proc Annu Conv Am Psychol Assoc* 1970:557–8.

Bouman TK, Luteijn F, Schoenmaker NA. Personality scales on the personal computer: an equivalency study. *Psycholoog* 1988;**23**:377–80.

Bratton GR, Newsted PR. Response effects and computer-administered questionnaires: the role of the entry task and previous computer experience. *Behav Inform Tech* 1995;**14**:300–12.

Broadhead WE, Leon AC, Weissman MM, Barrett JE, Blacklow RS, Gilbert TT, *et al.* Development and validation of the SDDS-PC screen for multiple mental disorders in primary care. see comment. *Arch Fam Med* 1995;**4**:211–19.

Brodey BB, Rosen CS, Brodey IS, Sheetz BM, Steinfeld RR, Gastfriend DR. Validation of the Addiction Severity Index (ASI) for internet and automated telephone self-report administration. *J Subst Abuse Treat* 2004;**26**:253–9.

Broome KM, Knight K, Joe GW, Simpson D. Evaluating the drug-abusing probationer: clinical interview versus self- administered assessment. *Crim Justice Behav* 1996;**23**:593–606.

Bruijnzeels MA, van der Wouden JC, Foets M, Prins A, van den Heuvel WJ. Validity and accuracy of interview and diary data on children's medical utilisation in The Netherlands. *J Epidemiol Community Health* 1998;**52**:65–9.

Brummett BH, Maynard KE, Babyak MA, Haney TL, Siegler IC, Helms MJ, *et al.* Measures of hostility as predictors of facial affect during social interaction: evidence for construct validity. *Ann Behav Med* 1998;**20**:168–73.

Butler SF, Budman SH, Fernandez K, Jamison RN. Validation of a screener and opioid assessment measure for patients with chronic pain. *Pain* 2004;**112**:65–75.

Calmels P, Vedel E, Bethoux F, Charmet E, Minaire P. The Functional Independance Measure (FIM). Interest of the telephone administration. French. *Ann Readapt Meas Phys* 1994;**37**:469–76.

Cameron E, Sinclair W, Tiplady B. Validity and sensitivity of a pen computer battery of performance tests. *J Psychopharmacol* 2001;**15**:105–10.

Carmelli D, Rosenman RH, Chesney MA. Stability of the type A structured interview and related questionnaires in a 10-year follow-up of an adult cohort of twins. *J Behav Med* 1987;**10**:513–25.

Carp FM, Carp A. The validity, reliability and generalizability of diary data. *Exp Aging Res* 1981;**7**:281–96.

Cartwright A. Interviews or postal questionnaires? Comparisons of data about women's experiences with maternity services. *Milbank Q* 1988;**66**:172–89.

Coderre F, Mathieu A, St-Laurent N. Comparison of the quality of qualitative data obtained through telephone, postal and email surveys. *Int J Market Res* 2004;**46**:347–57.

Craig CL, Marshall AL, Sjostrom M, Bauman AE, Booth ML, Ainsworth BE, *et al.* International physical activity questionnaire: 12-country reliability and validity. See comment. *Med Sci Sports Exerc* 2003;**35**:1381–95.

Crowell JA, Treboux D, Waters E. The Adult Attachment Interview and the Relationship Questionnaire: relations to reports of mothers and partners. *Pers Relat* 1999;**6**:18 January.

Damschroder LJ, Baron J, Hershey JC, Asch DA, Jepson C, Ubel PA. The validity of person tradeoff measurements: randomised trial of computer elicitation versus face-to-face interview. *Med Decis Making* 2004;**24**:170–80.

Dansky BS, Saladin ME, Brady KT, Kilpatrick DG, Resnick HS. Prevalence of victimisation and posttraumatic stress disorder among women with substance use disorders: comparison of telephone and in-person assessment samples. *In J Addicts* 1995;**30**:1079–99.

Davila J, Cobb RJ. Predicting change in self-reported and interviewer-assessed adult attachment: tests of the individual difference and life stress models of attachment change. *Pers Soc Psychol Bull* 2003;**29**:859–70.

De Leeuw ED, Mellenbergh GJ, Hox JJ. The influence of data collection method on structural models: a comparison of a mail, a telephone, and a face-to-face survey. *Socio Meth Res* 1996;**24**:443–72.

Dell'Osso L, Armani A, Rucci P, Frank E, Fagiolini A, Corretti G, *et al.* Measuring mood spectrum: comparison of interview (SCI-MOODS) and self- report (MOODS-SR) instruments. *Compr Psychiatr* 2002;**43**:69–73.

Derogatis LR. The Derogatis Interview for Sexual Functioning (DISF/DISF-SR): an introductory report. *J Sex Marital Ther* 1997;**23**:291–304.

Dickson JP, MacLachlan DL. Fax surveys: return patterns and comparison with mail surveys. *J Market Res* 1996;**33**:108–13.

Dimock PH, Cormier P. The effects of format differences and computer experience on performance and anxiety on a computer-administered test. *Meas Eval Couns Dev* 1991;**24**:119–26.

Donovan RJ, Holman CD, Corti B, Jalleh G. Face-to-face household interviews versus telephone interviews for health surveys. *Aust New Zeal J Publ Health* 1997;**21**:134–40.

Engle A, Lynn LL, Koury K, Boyar AP. Reproducibility and comparability of a computerised, self-administered food frequency questionnaire. *Nutrition and Cancer* 1990;**13**:281–92.

EPIC GoS. Relative validity and reproducibility of a diet history questionnaire in Spain. II. Nutrients. *Int J Epidemiol* 1997;**26**(Suppl. 1):S100–9.

Epstein JF, Barker PR, Kroutil LA. Mode effects in self-reported mental health data. *Publ Opin Q* 2001;**65**:529–49.

Falinower S, Martret P, Lombart B, Reti E, Krause D, Annequin D. [Self-report of acute pain by children using an electronic version of the faces pain scale – revised on the PalmOne personal data assistant.] *Douleurs* 2004;**5**:249–57.

Fechner-Bates S, Coyne JC, Schwenk TL. The relationship of self-reported distress to depressive disorders and other psychopathology. *J Consult Clin Psychol* 1994;**62**:550–9.

Feldman-Naim S, Myers FS, Clark CH, Turner EH, Leibenluft E. Agreement between face-to-face and telephone-administered mood ratings in patients with rapid cycling bipolar disorder. *Psychiatr Res* 1997;**71**:129–32.

Fernandez RR, Cruz JJ, Mata GV. Validation of a quality of life questionnaire for critically ill patients. see comment. *Intensive Care Medicine* 1996;**22**:1034–42.

Ferraroni M, Decarli A, Franceschi S, Vecchia CL, Enard L, Negri E, *et al.* Validity and reproducibility of alcohol consumption in Italy. *Int J Epidemiol* 1996;**25**:775–82.

Ferriter M. Computer aided interviewing in psychiatric social work. *Comput Hum Serv* 1993;**9**:59–66.

Fichter MM, Quadflieg N. Comparing self- and expert rating: a self-report screening version (SIAB-S) of the structured interview for anorexic and bulimic syndromes for DSM-IV and ICD-10 (SIAB-EX). *Eur Arch Psychiatr Clin Neurosci* 2000;**250**:175–85.

Fictenberg NL, Putnam SH, Mann NR, Zafonte RD, Millard AE. Insomnia screening in postacute traumatic brain injury:utility and validity of the Pittsburgh Sleep Quality Index. *Am J PMR* 2001;**80**:339–45.

Fink LA, Bernstein D, Handelsman L, Foote J, Lovejoy M. Initial reliability and validity of the childhood trauma interview: a new multidimensional measure of childhood interpersonal trauma. *Am J Psychiatr* 1995;**152:**1329–35.

Foa EB, Riggs DS, Dancu CV, Rothbaum BO. Reliability and validity of a brief instrument for assessing post-traumatic stress disorder. *J Trauma Stress* 1993;**6**:459–73.

Fournier L, Kovess V. A comparison of mail and telephone interview strategies for mental health surveys. *Can J Psychiatr Rev Canad Psychiatr* 1993;**38**:525–33.

Fowler FJJ, Gallagher PM, Nederend S. Comparing telephone and mail responses to the CAHPS survey instrument. Consumer Assessment of Health Plans Study. *Med Care* 1999;**37**(Suppl. 3):MS41–9.

Freeman TR, Stewart M, Birtwhistle R, Fisher DC. Health diaries for monitoring events following immunisation. *Can J Publ Health* 2000;**91**:426–30.

French CC, Beaumont JG. The reaction of psychiatric patients to computerised assessment. *Br J Clin Psychol* 1987;**26**:267–78.

Furbee PM, Sikora R, Williams JM, Derk SJ. Comparison of domestic violence screening methods: a pilot study. *Ann Emerg Med* 1998;**31**:495–501.

Galan I, Rodriguez-Artalejo F, Zorrilla B. Telephone versus face-to-face household interviews in the assessment of health behaviours and preventative practices. *Gac Sanit* 2004;**18**:440–50.

Gill L, Shand P, Fuggle P, Dugan B, Davies S. Pain assessment for children with sickle cell disease: improved validity of diary keeping versus interview ratings. *Br J Health Psychol* 1997;**2**:131–40.

Goetz SM, Stuck AE, Hirschi A, Gillmann G, Dapp U, Minder CE, *et al.* [A multidimensional questionnaire as a component of preventative geriatric assessment: comparison of self-assessment version with the interview version.] *Soz Praventivmed* 2000;**45**:134–46.

Gomez-Peresmitre G, Granados A, Jauregui J, Pineda Garcia G, Tafoya Ramos SA. Body image measurement: paper and pencil and computerised test versions. *Rev Mex Psicol* 2000;**17**:89–99.

Grey A, Jackson DN, Howard JH. Validation of the survey of work styles: a profile measure of the type A behaviour pattern. *J Clin Epidemiol* 1989;**42**:209–16.

Groves R. On the mode of administering a questionnaire and responses to open-end items. *Soc Sci Res* 1978;**7**:257–71.

Hanscom B, Lurie JD, Homa K, Weinstein JN. Computerised questionnaires and the quality of survey data. *Spine* 2002;**27**:1797–801.

Hansell S, Sparacino J, Ronchi D, Strodtbeck FL. Ego development responses in written questionnaires and telephone interviews. *J Pers Soc Psychol* 1984;**47**:1118–28.

Harrell TH, Honaker LM, Hetu M, Oberwager J. Computerised versus traditional administration of the Multidimensional Aptitude Battery-Verbal scale: an examination of reliability and validity. *Comput Hum Behav* 1987;**3**:129–37.

Hayes V, Morris J, Wolfe C, Morgan M. The SF-36 health survey questionnaire: is it suitable for use with older adults? see comment. *Age Ageing* 1995;**24**:120–5.

Hebert JR, Ebbeling CB, Matthews CE, Hurley TG, Ma Y, Druker S, *et al.* Systematic errors in middle-aged women's estimates of energy intake: comparing three self-report measures to total energy expenditure from doubly labelled water. *Ann Epidemiol* 2002;**12**:577–86.

Heithoff KA, Wiseman EJ. Reliability of paper-pencil assessment of drug use severity. *Am J Drug Alcohol Abuse* 1996;**22**:109–22.

Herman S, Blumenthal JA, Haney T, Williams RB, Barefoot J. Type As who think they are type Bs: discrepancies between self-ratings and interview ratings of the type A (coronary-prone) behaviour pattern. *Br J Med Psychol* 1986;**59**:83–8.

Herzog A, Rodgers WL, Kulka RA. Interviewing older adults: a comparison of telephone and face-to-face modalities. *Publ Opin Q* 1983;**47**:405–18.

Holbrook AL, Green MC, Krosnick JA. Telephone versus face-to-face interviewing of national probability samples with long questionnaires: comparisons of respondent satisficing and social desirability response bias. *Publ Opin Q* 2003;**67**:79–125.

Hollowell CM, Patel RV, Bales GT, Gerber GS. Internet and postal survey of endourologic practice patterns among American urologists. *J Urol* 2000;**163:**1779–82.

Holst E, Holstein BE. [Sociomedical survey among the elderly in 10 EEC countries. An analysis based on non-respondents to a questionnaire survey of the population 70–95 years of age living in 4 Danish communities.] *Ugeskr Laeger* 1990;**152:**225–7.

Hoppe MJ, Gillmore MR, Valadez DL, Civic D, Hartway J, Morrison DM. The relative costs and benefits of telephone interviews versus self-administered diaries for daily data collection. *Eval Rev* 2000;**24**:102–16.

Horton SV, Lovitt TC. A comparison of two methods of administering group reading inventories to diverse learners: computer versus pencil and paper. *RASE* 1994;**15**:378–90.

Howe A, Bath P, Goudie F, Lothian K, McKee K, Newton P, *et al.* Getting the questions right: an example of loss of validity during transfer of a brief screening approach for depression in the elderly. *Int J Geriatr Pyschiatr*2000;**15**:650–5.

Jain MG, Harrison L, Howe GR, Miller AB. Evaluation of a self-administered dietary questionnaire for use in a cohort study. *Am J Clin Nutr* 1982;**36**:931–5.

Janofsky A. Affective self-disclosure in telephone versus face to face interviews. *J Humanist Psychol* 1971;**11**:93–103.

Johnson TP, O'Rourke DP, Burris JE, Warnecke RB. An investigation of the effects of social desirability on the validity of self-reports of cancer screening behaviours. *Medical care* 2005;**43**:565–73.

Jones JR, McWilliam C. The Geriatric Mental State Schedule administered with the aid of a microcomputer: a pilot study. *Int J Geriatr Pyschiatr* 1989;**4**:215–19.

Karlberg L, Krakau I, Sjoden PO, Unden AL. Psychometric properties of a brief self-report Type A questionnaire for use in primary health care. *Scand J Prim Health* 1997;**15**:52–6.

Kartashov AI, Buzin VV, Glazunov IS, Abol'ian LV. Results of testing the basic questionnaire SINDI in the evaluation of prophylaxis program development. *Sov Zdravookh* 1**991**:45–9.

Kazarian SS, Malla AK, Cole JD, Baker B. Comparisons of two expressed emotion scales with the Camberwell Family Interview. *J Clin Psychol* 1990;**46**:306–9.

Khan MS, Chaliha C, Leskova L, Khullar V. The relationship between urinary symptom questionnaires and urodynamic diagnoses: an analysis of two methods of questionnaire administration. *BJOG* 2004;**111:**468–74.

Kim JS, Kim GJ, Lee JM, Lee CS, Oh JK. HAIS (Hanil Alcohol Insight Scale): validation of an insight-evaluation instrument for practical use in alcoholism. *J Stud Alcohol* 1998;**59**:52–5.

Kimball JC. Career Interest Search: a prototype, computer-assisted occupational interest inventory for functionally illiterate adults. *J Employ Counsel* 1988;**25**:180–85.

Kittel F, Kornitzer M, Zyzanski SJ, Jenkins CD, Rustin RM, Degre C. Two methods of assessing the type A coronary-prone behaviour pattern in Belgium. *J Chron Dis* 1978;**31**:147–55.

Knapp H, Kirk SA. Using pencil and paper, internet and touch-tone phones for self-administered surveys: does methodology matter? *Comput Hum Behav* 2003;**19**:117–34.

Kobak KA. A comparison of face-to-face and videoconference administration of the Hamilton Depression Rating Scale. *J Telemed Telecare* 2004;**10**:231–5.

Kobak KA, Greist JH, Jefferson JW, Mundt JC, Katzelnick DJ. Computerised assessment of depression and anxiety over the telephone using interactive voice response. *MD Computing* 1999;**16**:64–8.

Kobak KA, Reynolds WM, Greist JH. Development and validation of a computer-administered version of the Hamilton Rating Scale. *Psychol Assess* 1993;**5**:487–92.

Kobak KA, Reynolds WM, Rosenfeld R, Greist JH. Development and validation of a computer-administered version of the Hamilton Depression Rating Scale. *Psychol Assess* 1990;**2**:56–63.

Kobak KA, Taylor LH, Dottl SL, Greist JH, Jefferson JW, Burroughs D, *et al.* Computerised screening for psychiatric disorders in an outpatient community mental health clinic. *Psychiatr Serv* 1997;**48**:1048–57.

Koch WR, Dodd BG, Fitzpatrick SJ. Computerised adaptive measurements of attitudes. *Meas Eval Couns Dev* 1990;**23**:20–30.

Korner-Bitensky N, Wood-Dauphinee S. Barthel Index information elicited over the telephone. Is it reliable? *Am J PMR* 1995;**74**:9–18.

Koson D, Kitchen C, Kochen M, Stodolosky D. Psychological testing by computer: Effect on response bias. *Educ Psychol Meas* 1970;**30**:803–10.

Kreindler D, Levitt A, Woolridge N, Lumsden CJ. Portable mood mapping: the validity and reliability of analogue scale displays for mood assessment via hand-held computer. *Psychiatr Res* 2003;**120:**165–77.

Kumanyika SK, Mauger D, Mitchell DC, Phillips B, Wright-Smiciklas H, Palmer JR. Relative validity of food frequency questionnaire nutrient estimates in the Black Women's Health Study. *Ann Epidemiol* 2003;**13**:111–18.

Lawrence E, Heyman RE, Leary KDO. Correspondence between telephone and written assessments of physical violence in marriage. *Behav Ther* 1995;**26**:671–80.

Lee DT, Yip SK, Chiu HF, Leung TY, Chan KP, Chau IO, *et al.* Detecting postnatal depression in Chinese women. Validation of the Chinese version of the Edinburgh Postnatal Depression Scale. see comment. *Br J Psychiatr* 1998;**172:**433–7.

Leece P, Bhandari M, Sprague S, Swiontkowski MF, Schemitsch EH, Tornetta P, *et al.* Internet versus mailed questionnaires: a randomised comparison (2). *J Med Internet Res* 2004 Oct 29;**6**:e39; PMID:15631963. *J Med Internet Res* 2004;**6**:e30. [Erratum appears in *J Med Internet Res* 2004;**6**:e38.]

Lefkowitz J, Katz ML. Validity of exit interviews. *Person Psychol* 1969:445–55.

Lenert LA, Sherbourne CD, Reyna V. Utility elicitation using single-item questions compared with a computerised interview. *Med Decis Making* 2001;**21**:97–104.

Lewinsohn PM, Rohde P, Gau JM. Comparability of self-report checklist and interview data in the assessment of stressful life events in young adults. *Psychol Rep* 2003;**93**:459–71.

Lewis G. Assessing psychiatric disorder with a human interviewer or a computer. *J Epidemiol Community Health* 1994;**48**:207–10.

Lilford RJ, Bourne G, Chard T. Comparison of information obtainable by computerised and manual questionnaires in an antenatal clinic. *Med Informats* 1982;**7**:315–20.

Link MW, Mokdad A. Moving the behavioural risk factor surveillance system from RDD to multimode: a web/mail/telephone experiment. *Am Stat Assoc* 2005:3889–96.

Lipton RB, Stewart WC, Solomon S. Questionnaire versus clinical interview in the diagnosis of headache. *Headache* 1992;**32**:55–6.

Lorei TW, Gurel LEE. Comparison of ex-mental patient employment information obtained by mail and by interview. *J Counsel Psychol* 1967:458–61.

McCormick MC, Workman-Daniels K, Brooks-Gunn J, Peckham GJ. When you're only a phone call away: a comparison of the information in telephone and face-to-face interviews. see comment. *J Dev Behav Pediatr* 1993;**14**:250–5.

MacDougall JM, Dembroski TM, Musante L. The structured interview and questionnaire methods of assessing coronary-prone behaviour in male and female college students. *J Behav Med* 1979;**2**:71–83.

McIntyre LM, Butterfield MI, Nanda K, Parsey K, Stechuchak KM, McChesney AW, *et al.* Validation of a Trauma Questionnaire in veteran women. *J Gen Intern Med* 1999;**14**:186–9.

Makolkin VI, Abbakumov SA, Alliluev IG, Pereverzev VS, Eligulashvili MM. Value of interviews and questionnaires in the differential diagnosis of pain in cardiac region. *Sovetskaia Meditsina* 1983:67–71.

Malcolm R, Sturgis ET, Anton RF, Williams L. Computer-assisted diagnosis of alcoholism. *Comput Hum Serv* 1989;**5**:163–70.

Mallinson S. The Short-Form 36 and older people: some problems encountered when using postal administration. *J Epidemiol Community Health* 1998;**52**:324–8.

Marin G, Marin BV. A comparison of three interviewing approaches for studying sensitive topics with Hispanics. *Hispanic J Behav Sci* 1989;**11**:330–40.

Marshall M, Sumner W. Family practice clerkship encounters documented with structured phrases on paper and hand-held computer logs. *Proceedings/AMIA* 2000 *Annual Symposium*:547–50.

Maruff P, Wood S, Currie J, McArthur-Jackson C, Malone V, Benson E. Computer-administered visual analogue mood scales:rapid and valid assessment of mood in HIV positive individuals. *Psychol Rep* 1994;**74**:39–42.

Mayes BT, Sime WE, Ganster DC. Convergent validity of type A behaviour pattern scales and their ability to predict physiological responsiveness in a sample of female public employees. *J Behav Med* 1984;**7**:83–108.

Merten T. Conventional and computerised test administration: the visual retention test. *Z Differ Diagn Psychol* 1999;**20**:97–115.

Miles EW, King WC. Gender and administration mode effects when pencil-and-paper personality tests are computerised. *Educ Psychol Meas* 1998;**58**:68–76.

Mittman C, Barbela T, McCaw D, Pedersen E. The respiratory disease questionnaire: use of a self-administered version. *Arch Environ Health* 1979;**34**:151–7.

Morrison DM, Leigh BC, Gillmore MR. Daily data collection: A comparison of three methods. *J Sex Res* 1999;**36**:76–81.

Nagoshi C, Walter D, Muntaner C, Haertzen CA. Validation of the Tridimensional Personality Questionnaire in a sample of male drug users. *Pers Indiv Differ* 1992;**13**:401–09.

Neimeyer RA, Dingemans PM, Epting FR. Convergent validity, situational stability and meaningfulness of the Threat Index. *Omega-J Death Dying* 1977;**8**:251–65.

Nenad P, Lars-Goran O. Clinical Validation of the Swedish version of the quality of life inventory in crime victims with posttraumatic stress disorder and a nonclinical sample. *J Psychopathol Behav Assess* 2004;**26**:15–21.

Nordin S, Brämerson A, Murphy C, Bende M. A Scandinavian Adaptation of the Multi-Clinic Smell and Taste Questionnaire: Evaluation. *Acta OtoLaryngol* 2003;**123**:536–42.

Oliver C, McClintock K, Hall S, Smith M, Dagnan D, Stenfert-Kroese B. Assessing the severity of challenging behaviour: psychometric properties of the Challenging Behaviour Interview. *J Appl Res Intellect Disabil* 2003;**16**:53–61.

Palermo TM, Valenzuela D, Stork PP. A randomised trial of electronic versus paper pain diaries in children: impact on compliance, accuracy, and acceptability. *Pain* 2004;**107:**213–9.

Paykel ES, Prusoff BA, Klerman GL, DiMascio A. Self-report and clinical interview ratings in depression. *J Nerv Ment Dis* 1973;**156:**166–82.

Peck JR, Smith TW, Ward JR, Milano R. Disability and depression in rheumatoid arthritis. A multi-trait, multi-method investigation. *Arthritis Rheum* 1989;**32**:1100–6.

Pecoraro RE, Inui TS, Chen MS, Plorde DK, Heller JL. Validity and reliability of a self-administered health history questionnaire. *Public Health Reports* 1979;**94**:231–8.

Pederson LL, Baskerville JC, Ashley MJ, Lefcoe NM. Comparison of mail questionnaire and telephone interview as data gathering strategies in a survey of attitudes towards restrictions on cigarette smoking. *Can J Publ Health* 1985;**76**:179–82.

Persoons P, Luyckx K, Desloovere C, Vandenberghe J, Fischler B. Anxiety and mood disorders in otorhinolaryngology outpatients presenting with dizziness: validation of the self-administered PRIMEMD Patient Health Questionnaire and epidemiology. *Gen Hosp Psychiatr* 2003;**25**:316–23.

Peters ML, Sorbi MJ, Kruise DA, Kerssens JJ, Verhaak PF, Bensing JM. Electronic diary assessment of pain, disability and psychological adaptation in patients differing in duration of pain. *Pain* 2000;**84**:181–92.

Petrie K, Abell W. Responses of parasuicides to a computerised interview. *Comput Hum Behav* 1994;**10**:415–18.

Pettigrew LE, Wilson JT, Teasdale GM. Reliability of ratings on the Glasgow Outcome Scales from in-person and telephone structured interviews. *J Head Trauma Rehabil* 2003;**18**:252–8.

Piechowski MM, Miller NB. Assessing developmental potential in gifted children: a comparison of methods: correction. *Roeper Rev* 1995;**17**:238.

Polednak AP, Lane DS, Burg MA. Mail versus telephone surveys on mammography utilisation among women 50–75 years old. *Med Care* 1991;**29**:243–50.

Potts MK, Daniels M, Burnam MA, Wells KB. A structured interview version of the Hamilton Depression Rating Scale: evidence of reliability and versatility of administration. *J Psychiatr Res* 1990;**24**:335–50.

Quigley-Fernandez B, Tedeschi JT. The bogus pipeline as lie detector: two validity studies. *J Pers Soc Psychol* 1978;**36**:247–56.

Rabin JM, McNett J, Badlani GH. Computerised voiding diary. *Neurourol Urodyn* 1993;**12**:541–53;53–4.

Reich W, Earls F. Interviewing adolescents by telephone: is it a useful methodological strategy? *Compr Psychiatr* 1990;**31**:211–15.

Remschmidt H, Hirsch O, Mattejat F. [Reliability and validity of evaluation data collected by telephone.] *Z Kinder Jugenpsychiatr* 2003;**31**:35–49.

Reuband K, Blasius J. Face-to-face, telephone and mail surveys: response rates and patterns in a large city study. *Koln Z Soziol Sozialpsych* 1996;**48**:296–318.

Reuben DB, Valle LA, Hays RD, Siu AL. Measuring physical function in community-dwelling older persons: a comparison of self-administered, interviewer-administered, and performance-based measures. *J Am Geriatr Soc* 1995;**43**:17–23.

Richman-Hirsch WL, Olson-Buchanan JB, Drasgow F. Examining the impact of administration medium on examinee perceptions and attitudes. *J Appl Psychol* 2000;**85**:880–87.

Roccaforte WH, Burke WJ, Bayer BL, Wengel SP. Validation of a telephone version of the mini-mental state examination. *J Am Geriatr Soc* 1992;**40**:697–702.

Roger S, Gribble J, Turner C, Miller H. Entretiens autoadministrés sur ordinateur et mesure des comportements sensibles. *Population* 1999;**54**:231–50.

Rose G, McCartney P, Reid DD. Self-administration of a questionnaire on chest pain and intermittent claudication. *British J Prev Soc Med* 1977;**31**:42–8.

Rosenfeld P, Doherty LM, Vicino S, Kantor J. Attitude assessment in organisations: testing three microcomputer-based survey systems. *J Gen Psychol* 1989;**116:**145–54.

Rosenfeld R, Dar R, Anderson D, Kobak KA. A computer-administered version of the Yale-Brown Obsessive-Compulsive Scale. *Psychol Assess* 1992;**4**:329–32.

Ross SJ, Young CM. Mail versus Web questionnaires in municipal recreation settings: a comparative study of survey methodology. *Leisure/Loisir* 2002;**27**:115–35.

Rossier P, Wade DT. The Guy's Neurological Disability Scale in patients with multiple sclerosis: a clinical evaluation of its reliability and validity. *Clin Rehabil* 2002;**16**:75–95.

Rudiger K. Is computer assessment of obsession and compulsion applicable in obsessive–compulsive disorder? Preliminary results using the Hamburg Obsession Compulsion Inventory-Computer Short Form (HOCI-CS). *Comput Hum Behav* 1990;**6**:133–9.

Sarrazin MS, Hall JA, Richards C, Carswell C. A Comparison of computer-based versus pencil-and-paper assessment of drug use. *Res Soc Work Pract* 2002;**12**:669–83.

Sasyniuk NGH, Hollinshead TM, Hollinshead RL, Mohtadi RM. A prospective pilot study comparing paper based and computer based outcome instruments. Poster Session. *Clin J Sport Med* 2002;**12**:62.

Schaik PV, Ahmed T, Suvakovic N, Hindmarsh JR. Effect of an educational multimedia prostate program on the International Prostate Symptom Score. *Eur Urol* 1999;**36**:36–9.

Schuldberg D. Varieties of inconsistency across test occasions: effects of computerised test administration and repeated testing. *J Pers Assess* 1990;**55**:168–82.

Schumacher J, Hinz A, Hessel A, Brahler E. On the comparability of internet-based and paper-and-pencil surveys: a study with the Questionnaire of Recalled Parental Rearing Behaviour. *Diagnostica* 2002;**48**:172–80.

Schwartz NS, Mebane DL, Malony HN. Effects of alternative modes of administration on Rorschach performance of deaf adults. *J Pers Assess* 1990;**54**:671–83.

Schwenkmezger P, Hank P. Paper-and-pencil versus computer-assisted presentation of state–trait inventories: an equivalence test. *Diagnostica* 1993;**39**:189–210.

Searles JS, Helzer JE, Rose GL, Badger GJ. Concurrent and retrospective reports of alcohol consumption across 30, 90 and 366 days: interactive voice response compared with the timeline follow back. *J Stud Alcohol* 2002;**63**:352–62.

Sears RR. Comparison of interviews with questionnaires for measuring mothers' attitudes towards sex and aggression. *J Pers Soc Psychol* 1965;**2**:37–44.

Shaw RA, Crane J, Pearce N, Burgess CD, Bremner P, Woodman K, *et al.* Comparison of a video questionnaire with the IUATLD written questionnaire for measuring asthma prevalence. *Clin Exp Allergy* 1992;**22**:561–8.

Shrier LA, Shih MC, Beardslee WR. Affect and sexual behaviour in adolescents: a review of the literature and comparison of momentary sampling with diary and retrospective self-report methods of measurement. *Pediatrics* 2005;**115:**e573–81.

Siegel JM, Matthews KA, Leitch CJ. Validation of the type A interview assessment of adolescents: a multidimensional approach. *Psychosom Med* 1981;**43**:311–21.

Simon RJ, Fleiss JL, Fisher B, Gurland BJ. Two methods of psychiatric interviewing: telephone and face-to-face. *J Psychol* 1974;**88**:141–46.

Snyder-Ramos SA, Seintsch H, Bottiger BW, Motsch J, Martin E, Bauer M. Patient satisfaction and information gain after the preanesthetic visit: a comparison of face-to-face interview, brochure, and video. *Anaesth Analg* 2005;**100:**1753–8.

St-Pierre M, Beland Y. Mode Effects in the Canadian Community Health Survey: a comparison of CAPI and CATI. *Am Stat Assoc* 2**005:**4438–45.

Steketee G, Frost R, Bogart K. The Yale-Brown Obsessive Compulsive Scale: interview versus self-report. *Behav Res Ther* 1996;**34**:675–84.

Straus SG, Miles JA, Levesque LL. The effects of videoconference, telephone, and face-to-face media on interviewer and applicant judgments in employment interviews. *J Manag* 2001;**27**:363–81.

Stringfellow VL, Fowler FJJ, Clarridge BR. Evaluating mode effects on a survey of behavioural health care users. *Am Stat Assoc* 2**001**:1–5.

Stukenberg KW, Dura JR, Kiecolt-Glaser JK. Depression screening scale validation in an elderly, community-dwelling population. *Psychol Assess* 1990;**2**:134–8.

Sykes RC, Ito K. The effects of computer administration on scores and item parameter estimates of an IRT-based licensure examination. *Appl Psychol Meas*1997;**21**:51–63.

Tanofsky-Kraff M, Morgan CM, Yanovski SZ, Marmarosh C, Wilfley DE, Yanovski JA. Comparison of assessments of children's eating-disordered behaviours by interview and questionnaire. *Int J Eat Disord* 2003;**33**:213–24.

Thorson JA, Powell FC. Methodological note on use of the Centre for Epidemiological Studies Depression Scale with older samples. *Psychol Rep* 1999;**85**:823–4.

Tourangeau R, Smith TW. Asking sensitive questions: the impact of data collection mode, question format, and question context. *Publ Opin Q* 1996;**60**:275–304.

Tourangeau R, Couper M, Steiger DM. Humanising self-administered surveys: experiments on social presence in Web and IVR surveys. *Comput Hum Behav* 2003;**19**:24 January.

Van Den Kerkhof EG, Goldstein DH, Lane J, Rimmer MJ, Dijk JPV. Using a personal digital assistant enhances gathering of patient data on an acute pain management service: a pilot study. see comment. *Can J Anesth* 2003;**50**:368–75.

Van Den Kerkhof EG, Goldstein DH, Rimmer MJ, Tod DA, Lee HK. Evaluation of hand-held computers compared to pen and paper for documentation on an acute pain service. *Acute Pain* 2004;**6**:115–21.

Van Den Kerkhof EG, Goldstein DH, Blaine WC, Rimmer MJ. A comparison of paper with electronic patient-completed questionnaires in a preoperative clinic. *Anaesth Analg* 2005;**101:**1075–80.

Van de Vijver FJR, Harsveldt M. The incomplete equivalence of the paper-and-pencil and computerised versions of the General Aptitude Test Battery. *J Appl Psychol* 1994;**79**:852–59.

Van Der Zouwen J, De Leeuw ED. The Relationship between Mode of Administration and Quality of Data in Survey Research. *BMS* 1991;**31**:49–60.

Vansickle TR, Kimmel C, Kapes JT. Test–retest equivalency of the computer-based and paper-pencil versions of the Strong-Campbell Interest Inventory. *Meas Eval Couns Dev* 1989;**22**:88–93.

Verghese J, Katz MJ, Derby CA, Kuslansky G, Hall CB, Lipton RB. Reliability and validity of a telephone-based mobility assessment questionnaire. *Age Ageing* 2004;**33**:628–32.

Waters TJ. Further comparison of video tape and face-to-face interviewing. *Percept Mot Skills* 1975;**41**:743–6.

Weeks MF, Kulka RA, Lessler JT, Whitmore RW. Personal versus telephone surveys for collecting household health data at the local level. *Am J Public Health* 1983;**73**:1389–94.

Weiland SK, Mundt KA, Ruckmann A, Keil U. Self-reported wheezing and allergic rhinitis in children and traffic density on street of residence. *Ann Epidemiol* 1994;**4**:243–7.

Wiechmann D, Ryan AM. Reactions to computerised testing in selection contexts. *Int J Select Assess* 2003;**11**:215–29.

Wiedemann G, Rayki O, Feinstein E, Hahlweg K. The Family Questionnaire: development and validation of a new self-report scale for assessing expressed emotion. *Psychiatr Res* 2002;**109:**265–79.

Williams DA, Gendreau M, Hufford MR, Groner K, Gracely RH, Clauw DJ. Pain assessment in patients with fibromyalgia syndrome: a consideration of methods for clinical trials. *Clin J Pain* 2004;**20**:348–56.

Wilson JT, Edwards P, Fiddes H, Stewart E, Teasdale GM. Reliability of postal questionnaires for the Glasgow Outcome Scale. *J Neurotrauma* 2002;**19**:999–1005.

Wilson K, Roe B, Wright L. Telephone or face-to-face interviews?: a decision made on the basis of a pilot study. *Int J Nurs Stud* 1998;**35**:314–21.

Wong NY, Nenny S, Guy RJ, Seow-Choen F. Adults in a high-risk area are unaware of the importance of colorectal cancer: a telephone and mail survey. *Dis Colon Rectum*2002;**45**:946–50; quiz 51–4.

Wood M. Symptom assessment in epidemiology: a comparison between two methods. *Br J Soc Clin Psychol* 1971;**10**:355–59.

Wright L, May K, Jackson K. Exaggerated social control and its relationship to the type a behaviour pattern as measured by the structured interview. *J Res Pers* 1991;**25**:135–6.

Xia DY, Liao SS, He QY, Liao JF, Wang XC, Wu QH. A questionnaire-based survey on attitude and behaviour of sex among rural women in Hainan province. *Chin J Epidemiol* 2004;**25**:586–9.

Yates BT, Wagner JL, Suprenant LM. Recall of health-risky behaviours for the prior 2 or 4 weeks via computerised versus printed questionnaire. *Comput Hum Behav* 1997;**13**:83–110.

Zimmerman M, Pfohl B, Stangl D. Life events assessment of depressed patients: a comparison of self-report and interview formats. *J Hum Stress* 1986;**12**:13–19.

Appendix 7

Description of Minnesota Multiphasic Personality Inventory scales

The Minnesota Multiphasic Personality Inventory scales

Scale 1 — Hypochondriasis

This scale was originally developed to identify patients who manifested a pattern of symptoms associated with the label of hypochondria. All the items on this scale deal with subjects who are unrealistically concerned with bodily complaints. Scale 1 is designed to assess a neurotic concern over bodily functioning. A person who is actually physically ill will obtain only a moderate score on the hypochondriasis scale. These people will endorse their legitimate physical complaints, but will not endorse the entire range of vague physical complaints included in this scale.

Scale 2 — Depression

This scale focuses on lack of hope in the future, a general dissatisfaction with one's own life situation and poor morale. Low scores signify a general unhappiness with life, but high scores indicate clinical depression.

Scale 3 — Hysteria

This scale looks at hysterical reaction to stressful situations. People will often have a 'normal' facade and then break down when faced with high 'trigger' levels of stress. High scores on this scale indicate people that are more intelligent, better educated and from a higher social class. Women have predominantly scored higher than men on this scale.

Scale 4 — Psychopathic deviation

This scale measures social deviation and looks at lack of acceptance of authority and amorality. Higher scores on this scale are generally achieved by adolescents. This scale was originally developed to identify patients diagnosed as having a psychopathic personality. Scale 4 can be thought of as a measure of rebelliousness; a higher score will indicate rebellion and lower scores indicate an acceptance of authority.

Scale 5 — Masculinity–femininity

This scale was originally developed to identify homosexuality, but was unable to do so accurately. The masculinity–femininity scale is now used to measure how strongly an individual identifies with the traditional (pre-1960s) masculine or feminine role, intelligence, education and socioeconomic status.

Men on average tend to obtain higher scores on the masculinity–femininity scale. High scores are extremely uncommon among females. If a high score is achieved it can generally indicate a rejection of the traditional female role.

Scale 6 — Paranoia

This scale looks at paranoid symptoms such as suspiciousness, grandiose self-concepts, excessive sensitivity, ideas of reference, feelings of persecution and rigid opinions and attitudes. A high score on the paranoia scale indicates that the subject has strong, irrational suspicions and overestimates his or her own self-importance.

Scale 7 — Psychasthenia

This scale was originally designed to look at symptoms such as compulsion, obsessions, excessive doubt and unreasonable fears. Psychasthenia indicates conditions such as obsessive-compulsive disorder. The scale also highlights difficulties in concentration, self-criticism, abnormal fears and guilty feelings. High scores on the psychasthenia scale highlight that the subject may be tense and anxious and may have obsessive thoughts or compulsive behaviours.

Scale 8 — Schizophrenia

This scale assesses a wide variety of content areas, including bizarre thought processes and peculiar perceptions, social alienation, poor familial relationships, difficulties in concentration and impulse control, lack of deep interests, disturbing questions of self-worth and self-identity, and sexual difficulties. High scores on this scale indicate that the subject is withdrawn, may experience distortions of reality and can tend to act bizarrely.

Scale 9 — Hypomania

This scale tests for elevated mood, accelerated speech and motor activity, irritability, flight of ideas and brief periods of depression. A participant who achieves a high score on is likely to be outgoing, impulsive, overly active and excited.

Scale 0 — Social introversion

This scale looks at a person's inclination to withdraw from social contacts and responsibilities; thus, it will assess how shy or outgoing a person is. Hence, if a high score is achieved it indicates the subject is withdrawn, shy, inhibited and unassuming.

Validity scales

The authors also developed four validity scales to improve the overall accuracy of the measure, detect 'deviant test-taking attitudes' and gauge the accuracy of the other scales.

The 'cannot say' scale – ?

The 'cannot say' scale is the frequency of the number of items omitted or which have been marked both true and false on the whole outcome measure. If the scale has large number of missing items this can call into question the scores on all the other scales. The MMPI manual suggests that participants with 30 or more omitted items should be considered invalid and not interpreted. High scores on this scale can also indicate that the subject is indecisive.

The L scale

Originally called the 'lie' scale, this attempted to assess naive or unsophisticated attempts by people to present themselves in an overly favourable light. In terms of scoring, people who obtain high L scores are not willing to admit even minor shortcomings, hence, are deliberately trying to present themselves in a more positive way. People who are better educated and more sophisticated people from a high social class tend to score lower on the L scale.

The F scale

This is the deviant or rare response scale. The scale will analyse the items which are rarely endorsed by normal people. If less than 10% of the normal population sanction the item, but you endorse it, your F score would increase. For instance 'all laws should be eliminated'.

The F scale has three vital functions:

1. It is an index of test-taking attitude and is useful in detecting deviant response sets (i.e. faking good or faking bad).
2. If one can rule out profile invalidity, the F scale is a good indicator of degree of psychopathology, with higher scores suggesting greater psychopathology.
3. Scores on the F scale can be used to generate inferences about other characteristics and behaviours.

The K scale

The K scale was designed to analyse more subtle distortion of response, particularly clinically defensive response. The K scale was constructed by comparing the responses of a group of people who were known to be clinically deviant but who produced normal MMPI profiles with a group of normal people who produced normal MMPI profiles (no evidence of psychopathology in both). The K scale was subsequently used to alter scores on other MMPI scales. It was reasoned that people with high K values give scores on other scales which are too low, for instance if the participant achieves a high K score it will indicate that the subject is defensive and attempting to obscure symptoms. K is used to adjust the scores on other scales.

Appendix 8

Description of Short Form questionnaire-36 items health scales

Physical functioning

Measures how able a responder is to perform physical tasks without limitations due to health.

Role physical

Measures due to physical health, a responder has problems with work or other daily activities.

Bodily pain

Measures the severity and level of limitation due to bodily pain.

General health perception

Measures overall health.

Vitality

Measures energy levels and fatigue.

Social functioning

Measures the level of interference with social activities due to physical or emotional problems.

Role emotional

Measures how much emotional problems impact on work or daily activities.

Mental health

Measures levels of individual mental health.

Health Technology Assessment programme

Prioritisation Group

Members

HTA Commissioning Board

Members

HTA Commissioning Board *(continued)*

Observers

HTA Clinical Evaluation and Trials Board

Members

Observers

Diagnostic Technologies and Screening Panel

Members

Observers

Disease Prevention Panel

Members

Observers

External Devices and Physical Therapies Panel

Members

Chair,
Dr John Pounsford,
Consultant Physician North Bristol NHS Trust

Deputy Chair,
Professor E Andrea Nelson,
Reader in Wound Healing and Director of Research, University of Leeds

Professor Bipin Bhakta,
Charterhouse Professor in Rehabilitation Medicine, University of Leeds

Mrs Penny Calder,
Public contributor

Dr Dawn Carnes,
Senior Research Fellow, Barts and the London School of Medicine and Dentistry

Dr Emma Clark,
Clinician Scientist Fellow & Cons. Rheumatologist, University of Bristol

Mrs Anthea De Barton-Watson,
Public contributor

Professor Nadine Foster,
Professor of Musculoskeletal Health in Primary Care Arthritis Research, Keele University

Dr Shaheen Hamdy,
Clinical Senior Lecturer and Consultant Physician, University of Manchester

Professor Christine Norton,
Professor of Clinical Nursing Innovation, Bucks New University and Imperial College Healthcare NHS Trust

Dr Lorraine Pinnigton,
Associate Professor in Rehabilitation, University of Nottingham

Dr Kate Radford,
Senior Lecturer (Research), University of Central Lancashire

Mr Jim Reece,
Public contributor

Professor Maria Stokes,
Professor of Neuromusculoskeletal Rehabilitation, University of Southampton

Dr Pippa Tyrrell,
Senior Lecturer/Consultant, Salford Royal Foundation Hospitals' Trust and University of Manchester

Dr Nefyn Williams,
Clinical Senior Lecturer, Cardiff University

Observers

Dr Kay Pattison,
Senior NIHR Programme Manager, Department of Health

Dr Morven Roberts,
Clinical Trials Manager, Health Services and Public Health Services Board, Medical Research Council

Professor Tom Walley, CBE,
Director, NIHR HTA programme, Professor of Clinical Pharmacology, University of Liverpool

Dr Ursula Wells,
Principal Research Officer, Policy Research Programme, Department of Health

Interventional Procedures Panel

Members

Chair,
Professor Jonathan Michaels,
Professor of Vascular Surgery, University of Sheffield

Deputy Chair,
Mr Michael Thomas,
Consultant Colorectal Surgeon, Bristol Royal Infirmary

Mrs Isabel Boyer,
Public contributor

Mr Sankaran Chandra Sekharan,
Consultant Surgeon, Breast Surgery, Colchester Hospital University NHS Foundation Trust

Professor Nicholas Clarke,
Consultant Orthopaedic Surgeon, Southampton University Hospitals NHS Trust

Ms Leonie Cooke,
Public contributor

Mr Seumas Eckford,
Consultant in Obstetrics & Gynaecology, North Devon District Hospital

Professor Sam Eljamel,
Consultant Neurosurgeon, Ninewells Hospital and Medical School, Dundee

Dr Adele Fielding,
Senior Lecturer and Honorary Consultant in Haematology, University College London Medical School

Dr Matthew Hatton,
Consultant in Clinical Oncology, Sheffield Teaching Hospital Foundation Trust

Dr John Holden,
General Practitioner, Garswood Surgery, Wigan

Dr Fiona Lecky,
Senior Lecturer/Honorary Consultant in Emergency Medicine, University of Manchester/Salford Royal Hospitals NHS Foundation Trust

Dr Nadim Malik,
Consultant Cardiologist/Honorary Lecturer, University of Manchester

Mr Hisham Mehanna,
Consultant & Honorary Associate Professor, University Hospitals Coventry & Warwickshire NHS Trust

Dr Jane Montgomery,
Consultant in Anaesthetics and Critical Care, South Devon Healthcare NHS Foundation Trust

Professor Jon Moss,
Consultant Interventional Radiologist, North Glasgow Hospitals University NHS Trust

Dr Simon Padley,
Consultant Radiologist, Chelsea & Westminster Hospital

Dr Ashish Paul,
Medical Director, Bedfordshire PCT

Dr Sarah Purdy,
Consultant Senior Lecturer, University of Bristol

Dr Matthew Wilson,
Consultant Anaesthetist, Sheffield Teaching Hospitals NHS Foundation Trust

Professor Yit Chiun Yang,
Consultant Ophthalmologist, Royal Wolverhampton Hospitals NHS Trust

Observers

Dr Kay Pattison,
Senior NIHR Programme Manager, Department of Health

Dr Morven Roberts,
Clinical Trials Manager, Health Services and Public Health Services Board, Medical Research Council

Professor Tom Walley, CBE,
Director, NIHR HTA programme, Professor of Clinical Pharmacology, University of Liverpool

Dr Ursula Wells,
Principal Research Officer, Policy Research Programme, Department of Health

Pharmaceuticals Panel

Members

Observers

Psychological and Community Therapies Panel

Members

Observers

Feedback

The HTA programme and the authors would like to know your views about this report.

The Correspondence Page on the HTA website (www.hta.ac.uk) 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.

NETSCC, Health Technology Assessment
Alpha House
University of Southampton Science Park
Southampton SO16 7NS, UK
Email: hta@hta.ac.uk
www.hta.ac.uk

ISSN 1366-5278