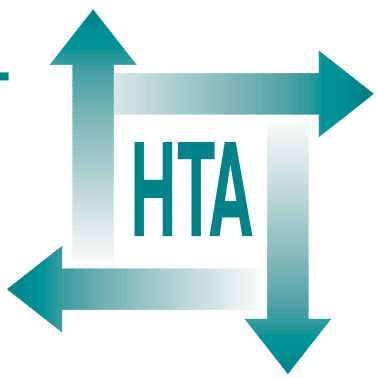
Health Technology Assessment 2001; Vol. 5: No. 5

Methodology

Eliciting public preferences for healthcare: a systematic review of techniques

M Ryan
DA Scott
C Reeves
A Bate
ER van Teijlingen
EM Russell
M Napper
CM Robb

Health Technology Assessment
NHS R&D HTA Programme



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

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

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

You can order HTA monographs from our Despatch Agents:

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

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

Contact details are as follows:

HTA Despatch
c/o Direct Mail Works Ltd
4 Oakwood Business Centre
Downley, HAVANT PO9 2NP, UK

Email: orders@hta.ac.uk
Tel: 02392 492 000
Fax: 02392 478 555
Fax from outside the UK: +44 2392 478 555

NHS libraries can subscribe free of charge. Public libraries can subscribe at a very reduced cost of £100 for each volume (normally comprising 30–40 titles). The commercial subscription rate is £300 per volume. Please see our website for details. Subscriptions can only be purchased for the current or forthcoming volume.

Payment methods

Paying by cheque
If you pay by cheque, the cheque must be in **pounds sterling**, made payable to *Direct Mail Works Ltd* and drawn on a bank with a UK address.

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

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

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

Please use the form on the HTA website (www.hta.ac.uk/htacd.htm). Or contact Direct Mail Works (see contact details above) by email, post, fax or phone. *HTA on CD* is currently free of charge worldwide.

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

Eliciting public preferences for healthcare: a systematic review of techniques

M Ryan[1*]
DA Scott[1]
C Reeves[1,2]
A Bate[1]
ER van Teijlingen[2]
EM Russell[2]
M Napper[1]
CM Robb[1]

[1] Health Economics Research Unit, University of Aberdeen, UK
[2] Department of Public Health, University of Aberdeen, UK

[*] Corresponding author

Competing interests: none declared

Published March 2001

This report should be referenced as follows:

Ryan M, Scott DA, Reeves C, Bate A, van Teijlingen ER, Russell EM, *et al*. Eliciting public preferences for healthcare: a systematic review of techniques. *Health Technol Assess* 2001;**5**(5).

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

NHS R&D HTA Programme

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

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

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

The research reported in this monograph was funded as project number 96/49/04.

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

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

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

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

ISSN 1366-5278

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

M

Contents

List of abbreviations

AHP	analytic hierarchy process
AIC	Akaike's information criterion*
CA	conjoint analysis
CE	closed-ended
CHC	community health council
EUT	expected utility theory
HCM	health council member
HE	health economist
HSR	health service researcher
IVF	*in vitro* fertilisation
MA	meta-analysis
MoV	measure of value
MS	medical sociologist
ns	not significant
OE	open-ended
PC	payment card
PEM	priority evaluator method
PGI	patient-generated index
PHC	public health consultant
PTO	person trade-off
QALY	quality-adjusted life-year
QDP	qualitative discriminant process
RPS	random paired scenarios
SDT	semantic differential technique
SEIQoL	schedule for the evaluation of individual quality of life
SEIQoL–DW	schedule for the evaluation of individual quality of life – direct weighting
SERVQUAL	service quality
SF-36	short-form 36
SG	standard gamble
TTO	time trade-off
VAS	visual analogue scale
WTA	willingness to accept
WTP	willingness to pay

* Used only in tables

Executive summary

Background

Limited resources coupled with unlimited demand for healthcare mean that decisions have to be made regarding the allocation of scarce resources across competing interventions. Policy documents have advocated the importance of public views as one such criterion. In principle, the elicitation of public values represents a big step forward. However, for the exercise to be worthwhile, useful information must be obtained that is scientifically defensible, whilst decision-makers must be able and willing to use it.

Aims and objectives

The aim was to identify techniques that could be reasonably used to elicit public views on the provision of healthcare. Hence, the objectives were:

- to identify research methods with the potential to take account of public views on the delivery of healthcare
- to identify criteria for assessing these methods
- to assess the methods identified according to the predefined criteria
- to assess the importance of public views *vis-à-vis* other criteria for setting priorities, as judged by a sample of decision-makers
- to make recommendations regarding the use of methods and future research.

Methods

A systematic literature review was carried out to identify methods for eliciting public views. Criteria currently used to evaluate such methods were identified. The methods identified were then evaluated according to predefined criteria.

A questionnaire-based survey assessed the relative importance of public views *vis-à-vis* five other criteria for setting priorities: potential health gain; evidence of clinical effectiveness; budgetary impact; equity of access and health status inequalities; and quality of service. Two techniques were used: choice-based conjoint analysis and allocation of points technique. The questionnaire was sent to 143 participants. A subsample was followed up with a telephone interview.

Results

The methods identified were classified as quantitative or qualitative.

Quantitative techniques

Quantitative techniques, classified as ranking, rating or choice-based approaches, were evaluated according to eight criteria: validity; reproducibility; internal consistency; acceptability to respondents; cost (financial and administrative); theoretical basis; whether the technique offered a constrained choice; and whether the technique provided a strength of preference measure.

Regarding ranking exercises, simple ranking exercises have proved popular, but their results are of limited use. The qualitative discriminant process has not been used to date in healthcare, but may be useful. Conjoint analysis ranking exercises did well against the above criteria.

A number of rating scales were identified. The visual analogue scale has proved popular within the quality-adjusted life-year paradigm, but lacks constrained choice and may not measure strength of preference. However, conjoint analysis rating scales performed well. Methods identified for eliciting attitudes include Likert scales, the semantic differential technique, and the Guttman scale. These methods provide useful information, but do not consider strength of preference or the importance of different components within a total score. Satisfaction surveys have been frequently used to elicit public opinion. Researchers should ensure that they construct sensitive techniques, despite their limited use, or else use generic techniques where validity has already been established. Service quality (SERVQUAL) appears to be a potentially useful technique and its application should be researched.

Three choice-based techniques with a limited application in healthcare are measure of value,

the analytical hierarchical process and the allocation of points technique, while those more widely used, and which did well against the pre-defined criteria, include standard gamble, time trade-off, discrete choice conjoint analysis and willingness to pay. Little methodological work is currently available on the person trade-off.

Qualitative techniques

Qualitative techniques were classified as either individual or group-based approaches. Individual approaches included one-to-one interviews, dyadic interviews, case study analyses, the Delphi technique and complaints procedures. Group-based methods included focus groups, concept mapping, citizens' juries, consensus panels, public meetings and nominal group techniques.

Six assessment criteria were identified: validity; reliability; generalisability; objectivity; acceptability to respondents; and cost.

Whilst all the methods have distinct strengths and weaknesses, there is a lot of ambiguity in the literature. Whether to use individual or group methods depends on the specific topic being discussed and the people being asked, but for both it is crucial that the interviewer/moderator remains as objective as possible. The most popular and widely used such methods were one-to-one interviews and focus groups. Both methods have potential problems with validity and reliability, and the researcher must minimise these problems at all stages of data collection, analysis and dissemination. The Delphi technique was widely used, but participants were only occasionally patients. It is proposed that the Delphi technique could be more widely used to gain patients' opinions. Citizens' juries were found to be very useful, especially with complex subject matter; the final decisions and opinions of participants are particularly valid and reliable because of the opportunity for deliberation. However, there are problems with generalisability as only very small numbers of people can be involved; they are also very time-consuming and therefore costly. Consensus panels are similar to citizens' juries but do not allow sufficient time for decision-making. They are less costly so cannot be dismissed. Public meetings, which are frequently used and are a quick and inexpensive way of gaining public opinion, are unrepresentative. Finally, complaints procedures only consider negative viewpoints, and therefore have limited value. Further work is needed to establish the appropriateness of case study analyses, concept mapping and nominal group techniques.

The importance of public views in priority setting

Both the choice-based conjoint analysis technique and the allocation of points method found the public's views to be important in the priority-setting exercise, although the relative rankings differed across the two techniques. In the follow-up telephone interviews, whilst the majority of respondents stated that the community had a role to play in decision-making, and that this role was (very) important in the context of priority setting, they ranked it as the least important of the six criteria.

Conclusions

Recommended techniques

There is no single, best method to gain public opinion. The method must be carefully chosen and rigorously carried out in order to accommodate the question being asked. Conjoint-based methods (including ranking, rating and choice-based), willingness to pay, standard gamble and time trade-off of the quantitative techniques and one-to-one interviews, focus groups, Delphi technique and citizens' juries of the qualitative methods are recommended. Likert scales, the semantic differential technique and Guttman scales are useful quantitative techniques for eliciting attitudes and knowledge.

Recommendations for future research

Researching techniques:

- the techniques recommended above should continue to be researched
- research to investigate analytical hierarchical process, measure of value, allocation of points, the qualitative discriminant process, SERVQUAL and person trade-off as quantitative methods with telephone, email and dyadic interview techniques, consensus panels, case study analyses, concept mapping and nominal group techniques as qualitative methods
- when addressing the above points, a priority area of research is to address the extent to which preferences for healthcare exist, as well as the cognitive strategies and decision-making heuristics respondents adopt when completing quantitative surveys. This should involve extensive qualitative work to inform the design and interpretation of quantitative studies.

General issues raised in the review:

- do the public want to be involved in healthcare decision-making?

- potential problems encountered with a preference for the *status quo*
- ethical issues in involving the public
- developments of frameworks to ensure public preferences are incorporated into priority setting.

Chapter 1

The purpose and plan of this review

Limited resources coupled with unlimited demand for healthcare mean that decisions have to be made regarding the allocation of scarce resources across competing interventions. Numerous criteria may be used to help this decision-making process. A combination of factors, including the rise of certain social movements (public, patients and feminist) and changes in the organisation of the NHS, has led to a growing awareness that the views of the public need to be taken into account. Policy documents agree.[1–4] For example, in its publication *Local Voices*,[1] the NHS Management Executive has advocated the need to take into account the views of local people when setting healthcare priorities.

In principle, attempts to involve public views in priority setting represents a big step forward. However, for the exercise to be worthwhile useful information must be obtained that is scientifically defensible and decision-makers must be willing to use it.

The **aim** of this project was to identify techniques that could be reasonably used to elicit public views in the provision of healthcare at both the micro and macro level.

Following this aim, the **objectives** were:

- to identify research techniques with the potential to take account of public views in the priority-setting exercise in healthcare
- to identify criteria for assessing these techniques
- to assess the techniques identified according to the predefined criteria
- to assess the importance of public views *vis-à-vis* other criteria for setting priorities, as judged by a sample of decision-makers
- to make recommendations regarding the use of techniques and future research.

The plan of this report is as follows. Chapter 2 sets out the methods used to address the first three objectives listed above. Following this, chapter 3 presents descriptions of the techniques identified in the review and chapter 4 identifies the criteria by which these techniques were assessed. Chapter 5 then presents the review of the quantitative techniques and chapter 6 the review of the qualitative techniques. Chapter 7 presents the methods and results of the primary research looking at the importance of public preferences in relation to other criteria for setting priorities (fourth objective above). Chapter 8 makes some overall conclusions and chapter 9 presents recommendations regarding the use of techniques and future research (the last objective above).

Chapter 2

Systematic review of techniques

Systematic reviews are "a scientific tool which can be used to summarise, appraise, and communicate the results and implications of otherwise unmanageable quantities of research".[5] The review presented here is concerned with methodological research, or, more specifically, with defining best practice in the area of eliciting public preferences in the provision of healthcare (as opposed to defining best medical practice). Methodological systematic reviews have become increasingly important in recent years,[6] leading to the question of what 'systematic' means for such reviews. What is clear is that methodological reviews have tended to take different approaches, both to identifying relevant material and to drawing conclusions from it. This has resulted from the broad range of issues addressed, with the nature of the review being very dependent on the question being addressed. Below we outline the methods used in our systematic methodological review to identify methods for eliciting public preferences in the delivery of healthcare. This was divided into three main parts: identifying the techniques; establishing the criteria by which to judge the methodological status of methods identified; and assessing the methods.

Identifying the techniques

Given the aim was to identify methods that had been used both within and outside health, a range of databases was searched. The core bibliographic databases used across all searches were MEDLINE (1966–), EMBASE (1980–), HealthSTAR (1975–) and the Social Science Citation Index (1981–). In addition, PsycLIT and the economics database EconLIT (1969–) was used extensively throughout. Other databases used were the Health Management Information Consortium Database (HMIC), CINAHL, Sociological Abstracts, and the Institute of Management International Databases (IMID). Search strategies were formulated using appropriate combinations of controlled vocabulary (where available) and free text terms. The search strategies are given in appendix 1.

Particular attention was paid to searching the grey literature, for example government publications and research reports through use of the Health Management Information Consortium Database, SIGLE (System for Grey Literature in Europe), IDEAS (Internet Documents in Economics Access Service), the Scottish Health Service Management Library's in-house database and the University of Aberdeen Health Economics Research Unit's own specialised library resource. The Health Service Management Centre's database, 'International Approaches to Priority Setting in Health Care' and the University of York's database of examples of consumer involvement in research were also consulted.

The Internet was searched using specific relevant websites and through a search engine (chiefly AltaVista®), where considered appropriate. The Social Science Information Gateway (SOSIG) was used to identify quality-assessed websites relevant to public opinion research. The Internet Documents in Economics Access Service was used to identify any relevant economics-based research reports.

Bibliographic searching was supplemented by citation searching of key articles, reviewing of references from key publications, and by consulting key texts.[7–12]

Following this **initial** search, the abstracts were scanned by the researchers (DAS for quantitative methods and CR for qualitative methods) and those that identified techniques that could be used for eliciting public preferences were obtained in hard copy. All other articles identified by the scope of the search strategy, for example those discussing the methodology of priority setting or patient involvement in decision-making, were excluded from the systematic review*. Papers not written in English were excluded. This resulted in a number of potentially useful techniques being identified which were examined in more detail. The authors then sought advice from a number

* Where such articles highlighted criteria and frameworks that are currently used for priority setting or general issues that are raised when eliciting public preferences, these were obtained in hard copy. The issues raised are covered in chapters 7 and 8 of this report.

of experts about whether any techniques could have been omitted from their list.

A **secondary** search was then conducted which focused on each individual technique. This consisted chiefly of straightforward searching across databases using the name of each technique as free text terms (unless available as controlled terms). For individual techniques, any acronyms, synonyms and any commonly used shorthand version of the name of the technique were included as free text terms. The search engine AltaVista was again used when searching for material relevant to a particular technique. An attempt was made to identify studies that had both applied the technique in healthcare, and had, either explicitly or implicitly, discussed methodological issues. Where the technique had been used in health, only those articles were considered but where the technique had not been used in health, the authors digressed further afield.

Following this secondary search, the abstracts of any additional papers were reviewed and if deemed relevant, a hard copy was obtained. No hand searching was undertaken within this secondary search. However, the reference lists of identified articles were searched for relevant studies and, in some cases, individual specialists in particular techniques were consulted for their reference lists. For example, one of the authors (MR) had considerable experience in the use of conjoint analysis (CA) and willingness to pay (WTP) and had colleagues who had also researched those areas. They were able to identify additional literature such as conference papers and works in progress. In addition, further searches were conducted on key individuals who were especially known for or who had contributed significantly to the adoption of a particular technique. For example, James Dolan had done a considerable amount of work on the analytic hierarchical process (AHP) and Eric Nord had also fronted much of the work on person trade-off (PTO). For techniques where little literature was found, for example Hoinville's priority evaluator method (PEM), articles citing the key papers were sourced. Any additional articles brought to light were obtained in hard copy.

All retrieved references were incorporated into a dedicated database using the bibliographic software package Reference Manager©.

Establishing the criteria by which to judge the methodological status of techniques identified

Although initially we did not discriminate between quantitative and qualitative methods for eliciting the views of the public, the literature search confirmed at an early point that methodological evaluation is at very different stages of formalisation for the two approaches†. Moreover, there is an ongoing debate about whether it is possible to apply the same criteria to both.[13] We therefore decided to handle them separately. We attempted to summarise the current debate on assessing quality in quantitative and qualitative research. To do this, articles identified in the systematic literature review described above were used, as well as additional literature the authors were familiar with. Following this, we defined the criteria by which the methods identified in the review were to be assessed.

Assessing the techniques

The quantitative and qualitative techniques identified in the searches were then reviewed according to predefined criteria. This formed the completion of the systematic review of the techniques. This project used only one researcher to review the techniques using the predefined criteria (although separate researchers were used for the quantitative and qualitative reviews). Whilst it is recognised that using more than one researcher to evaluate techniques would reduce researcher bias, time and resource constraints meant this was not possible. It may also be argued that, given the nature of the review (i.e. there were well-defined criteria with which to assess the techniques), interobserver reliability was not really a concern. However, this argument may apply more for the review of quantitative techniques than qualitative techniques. The synthesis of the results is qualitative, highlighting the strengths and weaknesses of the different approaches.

† Quantitative instruments allow estimation of 'numbers'. In contrast, qualitative instruments "begin with an intention to explore a particular area, collect 'data' and generate ideas and hypotheses from these data largely through **inductive reasoning**".[649] Qualitative research seeks to find answers to the 'why' questions as opposed to the 'how many' questions. Such methods help to find out what the 'meaning' of a certain phenomenon, event or relationship, for example, is for people and the 'context' in which these phenomena are experienced. For example, quantitative research may reveal that respondents hold certain attitudes, and qualitative research why they hold these attitudes.

Chapter 3

Techniques identified in the systematic review

Issues raised in the identification of techniques

Given the nature of the topic to be reviewed, it was apparent from the outset that a 'Cochrane-type' approach to identification of the relevant literature was not appropriate.[14] However, a systematic and comprehensive strategy was used. The **initial** strategy was designed to trawl the literature covering a number of different disciplines, both within and outwith the health-care field. This was reflected in the choice of bibliographic databases. Because the literature was an unknown quantity at the initial stage, the approach to formulating the search strategy was progressive rather than algorithmic. This approach is consistent with other methodological reviews.[6,14]

The search strategy was therefore developed (using a mix of controlled vocabulary and free text terms) to give high recall and therefore, by definition, low precision.[5] Consistent with other methodological reviews,[6,14,15] this approach resulted in a large number of articles for some techniques and a limited amount for others. We did not pursue all references, but made an attempt to stop the search when it appeared (from the abstracts) that the new literature emerging was of no additional benefit. Edwards and colleagues have referred to this truncation process as "theoretical saturation". They note that it is "important to cast the net wide by searching many types of literature to make sure that a particular line of argument was not missed, rather than to pursue every instance of the argument". Without such a strategy in this study it would not have been feasible to complete the project within the deadline.

It also became clear from the search strategy that whilst searching multiple databases lead to the identification of a wide range of techniques, it did so at the cost of duplication of references. However, this was deemed necessary by the research team if techniques were to be identified from a range of disciplines.

When searching for articles dealing with methodological issues such as validity/reliability, some databases included appropriate controlled thesaurus/vocabulary terms, for example in MEDLINE an appropriate MeSH term would be "Reproducibility of results". However, sometimes, a controlled vocabulary term was not available. Free text truncated terms were then used. This latter search technique, given that it searches for words regardless of context, resulted in a large amount of irrelevant references. However, this methodology was again deemed necessary by the research team.

Whilst the potential problems of a methodological systematic review were realised in this project, the project nevertheless resulted in information on a wide range of techniques that have been used, or may be used, to elicit public preferences in the delivery of healthcare. These techniques are described below. Most have been used to some extent in healthcare, though the frequency with which they have been applied varies widely*. Quantitative techniques are first described, followed by qualitative techniques.

Quantitative techniques for eliciting public views

The methods identified have been classified as ranking, rating or choice-based techniques. Appendix 2 presents a summary of studies identified (and methodological issues addressed; see chapter 5).

Ranking techniques

Simple ranking exercise

Ranking exercises in their simplest form ask respondents to give an ordinal ranking to

* Community health councils (CHCs) have been mentioned in the area of measuring public preferences.[650] Their remit is to consult with and represent the public, monitor services, and give advice and information.[651] However, CHCs are not a technique or instrument for obtaining public opinion, but a vehicle to aid the identification of concerns and attitudes of the public. Therefore, CHCs are not included in this review.

options[†]. Those options that achieve the highest ranking are viewed as the most important. Given their relative ease to analyse and complete by respondents, this approach has proved popular as a method of eliciting public preferences in healthcare.[16–27] The technique has been applied in a number of different contexts, including eliciting preferences for healthcare interventions, looking at the importance of patient characteristics in choices concerning who receives care, and valuing different aspects of clinical outcomes. For example, Groves[17] asked a random sample of the general public, managers and doctors in the UK to rank ten diverse healthcare interventions in order of priority for spending. Responses were averaged, and a ranking estimated for each of the ten interventions. Bowling and colleagues[16] adopted a slightly different ranking approach. In a study examining the attitudes towards priority setting in an inner London health district, the public and health professionals were asked to rank 16 health services in relation to their views on the needs of people in City and Hackney. The respondents were required to rank four services as 'essential', four as 'most important', four as 'important' and four as 'less important'. An overall priority score was estimated for each service/area/treatment by coding the possible responses from 1 (essential) to 4 (less important) and then averaging over these responses.

Qualitative discriminant process

The qualitative discriminant process (QDP) is a scoring and ranking process that has been developed in the business environment.[28–30] The technique is based on decision analysis techniques and is computer based. The distinguishing feature of this technique is that it involves moving from defining options in terms of **qualitative categories**, through to deriving a **numeric point estimate**, and finally solving a maximisation problem within given constraints. The technique involves three steps:

Step 1: the respondent is guided through three computer-based ranking rounds where they are asked to rank scenarios according to predefined qualitative descriptions. These three ranking stages are defined as broad, intermediate and narrow. Round 1 may involve defining scenarios as 'very high', 'high', 'average' and 'low'. This round will provide a preliminary ranking of alternatives. In round 2, respondents are required to further differentiate alternatives according to further subcategories, for example 'top', 'middle' and 'bottom'. The same process happens in round 3, with respondents now being asked to rank each alternative into further sub-subcategories, i.e. 'upper' and 'lower'. At each round, if two alternatives are defined equally, pairwise comparisons are used to determine the ranking. Following these three rounds, a ranking of alternatives is provided.

Step 2: this involves mapping the qualitative responses onto a numeric interval scale, and producing a vague real number for all options.

Step 3: linear programming techniques are used to identify the optimal solution within the given constraints.

Bryson and colleagues[28] provide an illustrative example within the context of identifying suitable candidates for the position of Dean of the Business School. The three criteria chosen to distinguish candidates were academic credential, management credentials and fundraising credentials. In this illustration, a multistage decision-making process was used, whereby the candidates were considered only if they had excellent academic credentials. The example provided relates to this stage. However, it should be noted that any stage could involve multiple criteria. In round 1, decision-makers were presented with a list of ten candidates and asked to rank them into the broad categories of 'excellent', 'very good', 'satisfactory' or 'poor'. Respondents are encouraged to move the candidates around until they are completely satisfied. The visual computer screens help this process. This stage yields a preliminary ranking of alternatives. Round 2 involves a second ranking exercise, whereby the decision-makers were asked to further distinguish the candidates into 'top', 'middle' and 'bottom'. Again, respondents are encouraged to move candidates around. Round 3 involves further

[†] Variations around such a simple ranking exercise were identified in the literature. **Plurality ranking** involves the respondent awarding their chosen option one point and the others zero. Options are summed for all respondents to achieve a rank order.[652] The **Borda measure**, derived from social choice theory, involves allocating zero points to the option ranked lowest, one point to the next lowest and so on until the highest ranked option receives the highest points. Scores are then weighted by numbers of respondents to give the Borda score for each alternative. This Borda score will take on a value between 0 and 1; the highest Borda score being the most desired option.[652,653] No studies were identified which had used this approach in healthcare.

differentiation into 'upper' and 'lower'. From these three rounds individual candidates are ranked. For example, the top candidate would be the one defined as 'excellent', 'top' and 'upper'. If candidates were defined equally then the decision-maker would have to make choices. Following the ranking exercise, a score for each candidate was assigned using an interval scale from 0 to 100, and defined as excellent (80–100), very good (60–80), satisfactory (30–60) and poor (0–30). Any subdivision of the ranking scale could be used.

The QDP is in the early stages of development, and there have been no applications eliciting public views in healthcare.

CA ranking exercises

CA (whether it adopts a ranking, rating or choice-based approach; see below for a description of the latter two) is rooted in Lancaster's theory of value.[31] This theory assumes that goods can be described by their characteristics, and that these goods enter an individual's benefit (utility) function as a combination, whereby the total utility gained from consuming a good is a function of the individual utilities from the characteristics of that good. The technique was developed in mathematical psychology,[32] and has been widely used in market research,[33] transport economics,[34] and environmental economics.[35,36] The technique is now gaining more widespread use in health economics to elicit public preferences for healthcare interventions. The technique involves five main stages:

Stage 1: identification of attributes, characteristics or criteria that are important in achieving the overall stated objective of the study.
Stage 2: assigning levels to these criteria.
Stage 3: using experimental designs to reduce the number of scenarios that individuals are presented with down to a manageable level.
Stage 4: eliciting preferences using ranking, rating or choice exercises.
Stage 5: analysing the data using regression technique.

From this it is possible to estimate: the marginal benefit or weight of individual criteria; the rate at which individuals trade between these criteria; and the overall benefit (or utility) scores for different combinations of levels of criteria. This latter score is achieved by multiplying the criterion weight by the level, and summing across all criteria for the given service.

A number of studies were identified that had used ranking-based CA exercises in healthcare.[21,37–45] Using the ranking approach, individuals are presented with a number of scenarios involving a different combination of criteria and levels, and asked to rank these. For example, Singh and colleagues[46] used the ranking CA technique to examine patient decision-making on growth hormone therapy. Six criteria were identified as important: risk of long-term side-effects (1:10,000 or 1:1,000,000); certainty of effect (50% or 100% of cases); amount of effects (1–2 inches or 4–5 inches in adult height); out-of-pocket cost ($100, $2000, or $10,000/year); route of treatment (daily injections or nasal spray)and child's attitude (likes or does not like therapy). Respondents were asked to provide a ranking of a number of scenarios that involved different combinations of the above criteria and levels. Further examples of the application of the ranking CA approach have been to determine physicians' decision-making process in the context of anti-infective drugs,[47] how to increase a hospital's patient population,[38] choosing alternative health plans,[21,39] alternative contraceptive methods,[40] evaluation of residents' clinical competence,[41] ambulatory care management,[43] and preferences for growth augmentation therapy.[46]

Rating techniques

Rating scales involve presenting individuals with criteria, scenarios or statements and asking them to respond with respect to their opinions, attitudes or knowledge on either a numerical or semantic scale. Numeric scales usually provide anchor descriptions such as 'best outcome' or 'worst outcome'. A large number of rating scale approaches were identified in the literature, with such scales being used to address a number of different issues, including estimating quality weights for health outcomes, understanding patient preferences for different aspects of treatment and assessing the public's attitudes and knowledge concerning different issues. Economists have tended to use rating scales to estimate utilities, whereas other social scientists have been more concerned with public attitudes. Amongst the latter a common application is satisfaction-type surveys. In this section we consider the range of such scales that have been used in healthcare to elicit public preferences.

Rating scales within the quality-adjusted life-year literature

Rating scales, usually referred to as visual analogue scales (VAS), have been widely used by economists to estimate utility weights within the quality-

adjusted life-year (QALY) literature.[48,49] QALYs were developed to take account of the fact that an individual may be concerned with the quality of their life as well as the quantity of life. To estimate QALYs, expected life years gained from given healthcare interventions are estimated and combined with information on the quality of these life years (via the estimation of utilities). For example, if a healthcare intervention results in a health state with a utility score of 0.85, and the individual would be in this health state for the remainder of life, say ten years, then the number of (undiscounted) QALYs would be 8.5. The QALYs gained from one healthcare intervention may be compared with QALYs obtained from alternative healthcare interventions, as well as from doing nothing. The QALY framework was developed to aid decisions concerning what healthcare interventions to provide‡.[50] Asking the public to estimate quality weights (as opposed to, for example, clinicians) will therefore incorporate public preferences into the decision-making process.

When using the VAS to estimate quality weights within the QALY framework, respondents are presented with a line with anchors at best imaginable health-state (with a score of 100) and worst imaginable health-state (with a score of 0). Such applications have often provided guidance to respondents by providing numbers at intervals. Respondents are then asked to indicate on this scale the point corresponding to either their own health-state, or another defined health-state. This point is taken as the quality weight for that health-state.

Rating scales within CA studies

Rating scales have been used in the CA literature to elicit patient and community preferences in healthcare. Here respondents are asked to rate scenarios on a numeric or semantic scale. The overall score is the dependent variable, and this is regressed against predefined levels of criteria.[51–65] Within healthcare this technique has been used to address a number of issues, including developing a handicap outcome measure,[58] establishing preferences for aspects of health services (including general practice,[60] scoliosis surgery,[51] antihistamine drugs,[59] dental health services[52]), obstetricians' referral patterns,[66] speciality selection by medical students,[54] design of an obstetrics unit,[56] factors influencing physicians' decision to operate,[57] physicians' weighting of clinical information[64] and hospital selection decisions.[65]

Schedule for the evaluation of individual quality of life

A variation of the CA rating technique, which shares the same theoretical foundations, is the schedule for the evaluation of individual quality of life (SEIQoL).[67] Using this technique respondents are first asked to identify the five most important areas of their life in terms of quality of life. The individual then rates each of these areas on a scale where the upper extreme is defined as 'as good as it could possibly be' (with a score of 100) and the lower extreme as 'as bad as it could possibly be' (with a score of 0). Computer-generated hypothetical cases are then presented to individuals which represent different combinations of the areas of life specified as important. Respondents are asked to rate these hypothetical cases on a scale with the same upper and lower extremes as defined above. Regression techniques are used to estimate the weights for the areas of life, and these weights are multiplied by the individual's self-rating and summed over the five areas to give a quality of life score. SEIQoL has been used to assess quality of life in hip replacement patients,[68] dementia patients,[67] the elderly[69] and sufferers of irritable bowel syndrome.[70]

Likert scale

Whilst economists have used rating scales to estimate quality weights or benefit scores, other social and behavioural scientists have tended to favour scales that are concerned with respondent's attitudes. A common technique used here is the Likert scale. This contains a series of opinion statements on a given issue.[9] Respondents' attitudes are elicited by presenting them with a series of statements and asking them their level of agreement on an agree–disagree continuous scale.[71–74] This is often an 'odd' number scale, with a neutral/undecided point in the middle. Likert scales have been used in healthcare to address a variety of issues including NHS managers' attitudes to capital charging,[75]

‡ Four techniques have been used in the literature to estimate quality weights: rating scales, magnitude estimation, standard gamble (SG) and time trade-off (TTO). SG and TTO are considered in the next section under choice-based techniques. Magnitude estimation is not considered in this report because it has been used infrequently; no more studies were identified beyond Brazier and colleagues' review,[253] and Brazier and colleagues did not recommend its use.

physicians' or nurses' views on the quality of healthcare,[76,77] dentists' and patients' judgements of ideal patients and dentists, respectively,[74,78,79] a patient's perceived quality of life,[80] views on priority setting and public involvement[73,81,82] and medical students' evaluations of patients.[83] Likert scales have also been extensively used in satisfaction studies (see below).

Semantic differential technique

Using the semantic differential technique (SDT), attitudes are elicited according to a number of opposite or polar adjectives at each end of a scale.[84–87] Responses are subsequently scored to give an overall 'attitude score'. The SDT, as well as being used in satisfaction scales (see below), has been used to assess attitudes regarding a number of factors in healthcare. These include attitudes towards: the menopause,[88,89] diabetes,[90] cervical screening,[84] health-related behaviour,[86,87] cancer detection methods,[91] HIV/AIDS[92] and schizophrenia.[93] For example, Swain and McNamara[92] carried out a study concerned with Irish pupils' attitudes to HIV/AIDS. The authors utilised the following bipolar adjectival eight-point scales: bad–good, false–true, painful–pleasurable, ugly–beautiful, unfair–fair, unimportant–important, unsafe–safe, worthless–valuable.

Guttman scales

Guttman scales have been applied in the fields of sociology, politics, psychology and consumer research to assess attitudes and knowledge.[71,94–97] They require respondents to respond to statements in terms of 'agree' or 'disagree'. Responses are coded as 'yes' for agree and 'no' for disagree. Scores of individual 'yes' responses are then estimated. From this it is possible to see which questions respondents as a group agree with.[96] Only five papers using the Guttman scale in healthcare were identified. These assessed attitudes towards: alcohol,[98] meat avoidance,[99] autopsy, organ donation and dissection,[100] long-stay care for the elderly[101] and dental health.[102]

Satisfaction surveys

Patient satisfaction surveys are a popular method for eliciting public opinion in healthcare, both in the UK and elsewhere.[73,103–158] These studies represent only a handful from the total number available. Bisset and Chesson[159] note that there have been over 4000 entries on MEDLINE over the past 5 years, as well as many others in the grey literature.

Fitzpatrick[144] outlines three reasons for conducting patient satisfaction studies: (i) an important outcome measure; (ii) useful in assessing consultations and patterns of communication; and (iii) used systematically, feedback enables choice between alternatives in organising or providing healthcare. This review focuses on the latter aspect of patient satisfaction studies. However, satisfaction studies are rarely designed and implemented purely for this third reason.

Within the UK, a number of patient satisfaction questionnaires have been developed with the aim of evaluating NHS services nationally. These have been concerned with different aspects of care, including hospital care,[160] general practice consultations,[146,147,161] out of hours primary medical care,[148] screening for breast cancer,[162] and maternity services.[163,164] However, most patient satisfaction surveys are based on questionnaires developed by the researchers themselves.[120]

Direct or indirect methods may be used to elicit patient satisfaction. The former method involves directly asking respondents their level of satisfaction with given aspects of care, with possible responses being on a rating scale ranging from 'very satisfied' to 'very dissatisfied'. Indirect methods involve inferring satisfaction indirectly from responses to such questions as 'the doctor provided me with all the information I wanted', with possible responses being on a Likert scale or on a Guttman scale (see pages 8 and 9).

SERVQUAL

Service quality (SERVQUAL) was developed in the marketing literature as a generic measure of service quality.[165,166] This technique is based in multiple discrepancy theory, and is concerned with the gap between expectations and perceptions. A Likert scale is used to measure 'quality' in terms of this difference. There are five dimensions to the SERVQUAL, and 22 statements referring to these dimensions. Respondents are first presented with 22 items concerning expectations about quality and asked to agree on a Likert-type scale, with 1 representing strongly disagree and 7 strongly agree. They are then presented with 22 items relating to their perceptions concerning the commodity being evaluated, and again asked to agree on a Likert-type scale, with 1 representing strongly disagree and 7 strongly agree. Quality is defined in terms of perceptions minus expectations. The results from such a survey help providers target those areas where the gap between expectations and perceptions is greatest. The technique has received much empirical attention in the healthcare literature.[166–182]

Choice-based techniques

Choice-based techniques force individuals to choose between alternatives presented to them. Given this, they all incorporate the choice criterion. Such techniques have taken a variety of forms when eliciting public preferences in healthcare.

Simple choice exercises

Choice exercises, at their simplest level, involve presenting respondents with scenarios that vary with respect to one characteristic, and asking them to choose between them. Charny and colleagues[183] asked respondents to choose between two hypothetical individuals differing by only one characteristic (age, marital status, gender, smoker/drinker status and employment). In a subset of this study, Lewis and Charny,[184] looking at the importance of age when allocating scarce healthcare resources, presented individuals with three scenarios. Each scenario forced the respondent to choose between two patients with leukaemia who were alike in all respects other than their age. They were told that only one of them could have treatment. A similar approach was adopted by Mooney and colleagues[185] in examining preferences for allocating healthcare gains. Here respondents were presented with six choices where equal health gains were to be allocated to different population groups based upon their age, sex, current health, socio-economic status, timing when benefits would be achieved and the number of individuals who would benefit.

An extension of this technique is random paired scenarios (RPS). This technique also involves presenting individuals with paired scenarios and asking them to make choices. However, more characteristics are included in the scenarios. Cross-tabulations are used to determine how often each attribute is selected as a 'winner'. The higher the selection rate, the more important that attribute. One study was identified which applied this technique in healthcare. This was concerned with the importance of different characteristics of patients when deciding on the allocation of scarce healthcare resources.[186] Characteristics identified as important were age (child, old patient), income (poor, rich patient), severity of disease (mild, severe), prognosis (good, poor), social status (low, high), cost of treatment (inexpensive, expensive) and origin of disease (self-acquired or not). In all, 24 hypothetical patients were created, each one containing three different characteristics of patients randomly selected from the list of six (e.g. child, rich patient, severe disease). Each respondent was presented with 12 random pairs. For each pair they were asked which one they would choose to be treated if only one could be paid for by society. Cross-tabulations were used to calculate the number of times each characteristic was selected as a winner. A selection rate of over 50% (meaning this characteristic was chosen as a winner in 50% of all the scenarios where it appeared) was defined as meaning that characteristic should be prioritised; 50% implied a neutral attitude; and less than 50% a negative prioritisation.

CA choice-based questions

Choice-based questions have been widely used within the CA framework in healthcare. Here individuals are presented with choices that involve different combinations of a good or service, and, for each choice, they state which they would choose or prefer. Possible responses may be either discrete (i.e. prefer A or prefer B)[187–203] or graded (i.e. strongly prefer A, prefer A, indifferent, prefer B, strongly prefer B).[204–209] Choice-based CA is gaining widespread use in healthcare and has been applied in a number of areas including: eliciting patient/community preferences in the delivery of health services,[9,187,191,196,197,210] establishing a consultant's preferences in priority setting,[211] evaluating health-states,[212,213] determining optimal treatments for patients,[64] evaluating alternatives within randomised controlled trials (RCTs)[199] and establishing patient preferences in the doctor–patient relationship.[202,203]

Analytic hierarchy process

The AHP has been widely used in the areas of social science, engineering and business to assess the relative importance (weights) of different criteria in the provision of a good or service, and following on from this, to derive scores for given goods and services.[214–224] There are four main stages involved in conducting an AHP study:[220]

Step 1 – Structure the problem: the first stage when using AHP is to construct a decision hierarchy for the overall problem being considered. This will usually involve stating the objective (level 1), the criteria that are important in achieving this objective (level 2), and the alternatives that may be used to achieve the objective, which will comprise different components of the criteria (level 3).

Step 2 – Making choices: the second step involves making relative judgements across

adjacent hierarchical levels. This is done with pairwise comparisons. First, the criteria identified (level 2) are assessed relative to the overall objective (level 1), and then the alternatives (level 3) are assessed in terms of each criterion (level 2). These choices are made in terms of preferences or importance on a nine-point 'intensity of importance' scale developed by Saaty.[223] This scale has definitions and explanations attached to the numeric scale, and is held to be on a ratio scale. For example, if criteria i and j from level 2 are compared in terms of their importance in meeting the objective, and criteria i is considered to be five times more important than option j, then j is 1/5 as important as i. Assuming n criteria at level 2, then $(n(n-1))/2$ are required. A pairwise matrix is created, comparing all criteria with each other. Following this, the alternatives (from level 3) are each considered in relation to each of the criteria (from level 2).

Step 3 – Estimating weights: following this, computer software is used to estimate weights for both the individual criteria in terms of meeting the overall objective, and the alternatives in terms of meeting the criteria.

Step 4 – Synthesising weights to score alternatives: following stage 3, a composite score is estimated for each alternative. This is made up by multiplying the relative weight of each criterion (from level 2) by the weight for each criterion within each alternative, and adding the results.

A number of studies were identified which had applied this technique to healthcare, although all were within the context of the American healthcare system[§]. Whilst the majority of studies have addressed issues relating to clinical decision-making and optimal treatment paths for patients,[225–237] preferences for alternative healthcare services[238–242] and systems[240,243,244] had also been addressed.

Standard gamble

SG is a choice-based technique that has been widely used by economists within the QALY paradigm. Using SG, a choice is presented which requires the respondent to choose between a certain outcome and a gamble. If the gamble is chosen, it may result in either a better outcome (with a probability p) or a worse outcome than the original (with a probability $1 - p$). The utility weight is gained through adjusting the probability of the best outcome until the subject is indifferent between the certain intermediate outcome and the gamble.[49,245,246] Given the known difficulties of answering a SG-type question, the technique is carried out with visual aids and via face-to-face interviews.

Time trade-off

Given the known difficulties of answering SG questions, the TTO technique was declared to estimate utility weights within the QALY framework.[247] This approach involves presenting individuals with a choice between living for a period t in a specified but less than perfect state (outcome B) versus having a healthier life (outcome A) for a time period h where $h < t$. Time h is varied until the respondent is in-different between the alternatives. The utility weight given to the less than perfect state is then h/t.

Person trade-off

The PTO technique may be seen as an extension to the QALY approach. Whilst the QALY approach values the health effects of interventions (obtained from VAS, SG or TTO), the PTO extends this to allow for distributive issues, i.e. who to treat.[248–250] The social value of a given healthcare intervention is derived by multiplying the utility gain from a given healthcare intervention (estimated using either VAS, SG or TTO) by both a social weight (SW), which is determined by the severity of the initial condition, and a potential for health weight (PW). Both social weight and potential for health weight are estimated using the PTO method. This involves asking individuals how many outcomes of one kind (x) are equivalent in social value to y outcomes of another kind. These descriptions will vary with respect to the distributive aspects that are considered important, i.e. severity and potential for health, although others may be added. So, a question may ask 'if there are x people in adverse health situation A and y people in adverse health situation B, and if you can only help (cure) one group, which group would you choose?',[250–253] x and y are varied until the respondent is indifferent between the two. The two states are then compared with each

[§] The search for AHP healthcare articles utilised the strategy detailed in chapter 2 in addition to the Hierarchon database of AHP applications on the Internet (www.expertchoice.com/hierarchon), and contact was made with James Dolan of the Unity Health System in Rochester, NY, USA.

other in terms of undesirability: B is x/y times as undesirable as A.[252,254,255]

Willingness to pay

WTP is a choice-based approach where individuals are presented with a choice between not having the commodity being valued, and having the commodity but forgoing a certain amount of money. The money that they are willing to forgo to have the commodity is their WTP for that commodity. Using survey techniques, WTP can be estimated using four techniques: open-ended (OE); bidding game; payment card (PC); and closed-ended (CE). Using the OE technique respondents are asked directly what the maximum amount of money is that they would be prepared to pay for a commodity. If the WTP study is carried out via an interview, the bidding technique can be used. Here individuals are asked if they would be willing to pay a specified amount. If they answer 'yes', the interviewer increases the bid until they reach amounts that the respondents are not willing to pay. If they answer 'no', the interviewer lowers the bid until they say 'yes'. WTP is estimated directly from the data provided. A variation on this is the PC technique. Here respondents are presented with a range of bids and asked to circle the amount that represents the most they would be willing to pay. An individual's true maximum WTP will lie somewhere in the interval between the circled amount and the next highest option. The CE approach asks individuals whether they would pay a specified amount for a given commodity, with possible responses being 'yes' or 'no'. The bid amount is varied across respondents and the only information obtained from each individual respondent is whether his or her maximum WTP is above or below the bid offered.

WTP has been widely used in healthcare to elicit public views.[256–258] The application of the WTP technique to healthcare has focused on using the PC and CE approaches.[259–263]

Measure of value

The measure of value (MoV) technique was developed by Churchman and Ackoff[264] within the context of optimal decision-making generally. Only one study was identified which had applied this technique to healthcare.[265] The method identifies the optimal bundle of services to be provided within the given resource constraints. Respondents are first asked to rank options. Suppose there are four options ranked as $O_1 > O_2 > O_3 > O_4$ (where > indicates preferred to). The respondent then assigns values ($v_1 \ldots v_4$) between 0 and 1 to each option. It is common to assign a value of 1.00 to the most preferred option, and successive values to the other options which may reflect strength of preferences. Assume that $O_1 = 1.00$, $O_2 = 0.80$, $O_3 = 0.50$ and $O_4 = 0.30$. These values are 'first estimates'. The respondent now compares O_1 against the combination $O_2 + O_3 + O_4$. Assuming he or she still prefers O_1, the values attached to each option (i.e. v_1, v_2, v_3, v_4) should be adjusted (if necessary) such that $v_1 > v_2 + v_3 + v_4$. Given the above values, $v_2 + v_3 + v_4 = 1.60 > v_1$, the values must be adjusted, for example: $v_2 = 0.4$, $v_3 = 0.25$ and $v_4 = 0.15$. The relative values of O_2, O_3 and O_4 must be retained. O_2 is then compared to $O_3 + O_4$. If $O_3 > O_4$, another adjustment of values is necessary since $v_3 + v_4 = 0.40 = v_2$. Adjusting v_3 to 0.30 and v_4 to 0.20 results in $v_3 + v_4 > v_2$. The final values $v_1 = 1.00$, $v_2 = 0.40$, $v_3 = 0.30$ and $v_4 = 0.20$ are then standardised by dividing them by the sum of $v_1 + v_2 + v_3 + v_4$. The resulting values $v_1 = 0.53$, $v_2 = 0.21$, $v_3 = 0.16$ and $v_4 = 0.10$ represent the values or scores for each option.[265,266] These scores are linked up with costs, and the optimal bundle or package is chosen from the resources available.

Allocation of points

A number of techniques have been developed that involve allocating points between alternative options or criteria. Such techniques assume that individuals know the weights they attach to different criteria, and can state them. This is in contrast with what has been called 'policy capturing' approaches, where it is assumed that individuals can provide overall evaluations, but that they cannot directly estimate weights for individual criteria. (Examples of 'policy capturing' approaches include all CA exercises introduced above, as well as the various methods for estimating QALY weights.) One of the earliest applications of this approach was the Hoinville PEM.[267,268] This technique was developed to aid town planners in taking account of the preferences of potential residents. Potential residents were allocated a number of points and asked to allocate them to the factors identified as important.

Variations of this method have been used in healthcare when asking patients to prioritise aspects of their lives that are most affected by a particular condition. Most of them are designed to assess change over time and therefore are constructed differently from measures intended to capture a single assessment. Importantly, all leave it to the respondent to identify the factors to be prioritised. One example of this is the patient-generated index (PGI). Here respondents are first asked to state the five most important

areas of their life affected by their condition. A sixth area is then defined as 'all other aspects of your life not mentioned above'. Respondents are then asked to rate how badly affected they are by their condition on a scale from 0 to 100, where 0 represents the worst possible situation and 100 the best. Respondents are then given 60 points and asked to spend them on improving the six aspects of their life. The points they give to the individual areas of their lives represent the relative importance of these areas. The individual proportions allocated to each area are multiplied by the rating to establish a score for an individual patient's quality of life.[269–273] Patient-generated quality of life measures are therefore estimated.

A similar approach was adopted in the development of the schedule for evaluation of individual quality of life – direct weighting (SEIQoL–DW) technique.[274,275] The SEIQoL–DW is an extension of the SEIQoL (introduced above) with a more direct method for estimating weights. Rather than using regression techniques to estimate weights indirectly, an allocation of points type approach is used to estimate weights directly. Respondents are presented with a pie chart with a 0–100 scale round it. Colour segments represent the different aspects of life, and respondents are asked to adjust the colours on the chart until each colour reflects the relative importance of that area. Quality of life scores are estimated in the same way as for the PGI.

A similar approach has been adopted in the priority-setting literature, although the technique has been referred to as the budget pie[#]. This approach differs from the PGI and SEIQoL–DW in that respondents are told the areas in which they are to allocate points.[276–278] This technique, and variations of it, have been applied in a number of settings.[276,277,279–286] Initial applications consisted of presenting an individual with a circle representing the budget for public services and asking respondents to cut the pie into pieces representing the amount for various budget items. Individuals were given examples of present budget allocations and were given examples of how the budget translated into service level provision.[277] Within healthcare, Honigsbaum and colleagues,[280] in a study concerned with setting priorities in Southampton Health Authority, asked 12 commissioners to allocate 100 points between five criteria used as a basis for priority setting: health gain; equity; local access; personal responsibility; and choice. A similar approach was applied in Oregon's attempt to set priorities for health services.[285,286] Here commissioners divided 100 points between three criteria: value to society; value to an individual at risk of needing the service; and essential to a healthcare package. Seventeen categories of health were then scored against these three criteria on a scale of 1–10.

A variation of the allocation of points approach was used by Ratcliffe[281] within a CA framework. The application was public preferences concerning the allocation of donor liver grafts. Respondents were presented with eight choices. Each choice presented individuals with two groups of patients that differed with respect to their age (40, 50 or 60), whether or not they had a alcohol-related liver disease, time already spent on the waiting list (3, 6 or 12 months) and whether they had previously had a transplant. For each of the eight choices respondents were asked to allocate 100 livers between the two groups of patients.

Qualitative techniques for eliciting public views

This section concentrates on qualitative techniques that have been, or could be, used to elicit public opinion in healthcare. The methods identified have been classified as individual or group-based approaches. Individual approaches included one-to-one interviews and the Delphi technique, and group-based included focus groups, citizens' juries, consensus panels and public meetings. Appendix 3 presents a summary of studies identified (and methodological issues addressed; see chapter 6).

Individual approaches

One-to-one interviews

The one-to-one interview is a technique based on a researcher (the interviewer) meeting a respondent (the interviewee) on an individual basis in order to seek the views of the latter. Interviews can be separated into three broad categories: structured, semi-structured and unstructured. While structured interviewing is very much a quantitative method,[9] the latter two are qualitative. Interviews can be conducted face-to-face, by telephone, or more recently by email.[287] Interviews have been used to investigate

[#] The budget pie has also been referred to as constant sum measurement, point voting system,[562,652] coupon scale and method of marks.[561]

a variety of topics within the healthcare setting. For example, Ayanian and colleagues[288] looked at the effect of racial differences on patients' preferences and expectations concerning renal transplantation; Dicker and Armstrong[289] investigated the underlying assumptions of interviewee responses on patients' views on priority setting; Williams and colleagues[290] used individual interviews as a way of establishing whether and how people evaluate services they receive; Crabtree and Miller[291] discussed the long interview as utilised in a study on pain perception, Wilson and colleagues[292] discussed the merits and pitfalls of using face-to-face and telephone interviews; Groves and Khan (in Baker[8]) investigated the costs of interviewing; Foster[287] looked at the possibility of using email services to conduct interviews; and Sohier[293] considered dyadic interviews.

The dyadic interview

Sohier[293] proposed the dyadic interview as a method where people may find it more acceptable to talk openly about difficult topics. The dyadic interview is a tool that enables two people to be interviewed at the same time. The two people are usually related to each other, for example a spouse or parent, or are close friends.

Case study analysis

Case studies are gained using standard unstructured interviews and observations, but the way the results are presented is unique. A description, along with the researcher's approach to understanding the case, is presented and descriptions of the major components of the case are detailed.[9]

The Delphi technique

The Delphi technique obtains attitudes and beliefs about a certain topic from members of a selected panel without the need for them to attend a meeting. Questionnaires or interview schedules are posted out to individuals; the questions are OE and seek feedback on an individual level. The information gathered is compiled into another questionnaire and sent back to the participating individuals who are asked to state their level of agreement with the statements made. This process is repeated and the rankings analysed to ascertain the degree of consensus.[9] Thus, there are both quantitative and qualitative elements to this method. Examples of where this method has been used in healthcare include: Guest and colleagues[294] who used "interviews with a Delphi panel" to gather information on a cost analysis of palliative care for terminally ill cancer patients; Endacott and colleagues[295] used a modified version of the Delphi technique to gain the opinions of paediatric intensive care sisters on the needs of critically ill children, and additionally evaluated the method; in Harrington's study,[296] senior practitioners were recruited to decide on research priorities on occupational health; Charlton and colleagues[297] looked at the spending priorities of different health professionals; Roberts and colleagues[298] compared the responses of consultant geriatricians and patients concerning important performance measures; Hadorn and Holmes[299] used Delphi to ask surgeons, clinicians and "relevant specialists" their opinions on elective surgical procedures with a view to formulating a set of criteria to assess the extent of expected benefit; Gabbay and Francis[300] applied the Delphi technique to ask general surgeons and anaesthetists on a national and local level their views on the maximum potential for day surgery; Wilson and Kerr[301] asked bioethics society members and their designates about important social values related to healthcare; Thomson and Ponder[302] used the Delphi technique to develop a survey technique; Gallagher and colleagues[303] and Burns and colleagues[304] involved patients in the Delphi process; and finally, Kastein and colleagues[305] dealt specifically with the issue of reliability of the technique.

Complaints procedures

Complaints procedures are noted here only because *Local Voices* mentioned this as a method of incorporating public preferences.[1] Using this approach, a system is set in place where users of healthcare can register their complaints. These complaints can then be investigated. Hull Health Authority has been cited as using such an approach.[1]

Group-based approaches

Focus groups

In a focus group a small number of individuals are brought together and encouraged to discuss interactively with other group members, under the guidance of a moderator or facilitator, a number of specified issues or topics. The interaction is a crucial part of the process. This technique is commonly used in the field of business studies to look at ways in which the problems and opportunities of marketing can be examined.[306] It is only in the last decade that it has been used in the social sciences,[307] and is now being used frequently in health research.[308] A focus group should be small, usually 8–12 people,[306] so as not to be intimidating and so that everybody has the opportunity to express their views. Traditionally there are several rules associated with

carrying out focus groups. It is stated by Morgan[309] that it is inappropriate to explore complex and sensitive issues, i.e. that the topic of discussion should be acceptable to participants, and that group members should not know one another before the interview commences.[310] Additionally, it is the task of the moderator to set ground rules and to ensure that participants remain focused on the discussion issue.[309,310] There are many examples available demonstrating the use of focus groups in the healthcare setting, and a selection of these are presented. Two studies that have been used to look specifically at priority-setting issues were the Somerset Health Authority[311] and also Kuder and Roeder.[312] Cohen and Garrett[313] looked at the sensitive subject of client/worker relationships in the residential mental health setting. Kitzinger[314] investigated people's attitude towards AIDS, whilst Ward and colleagues[315] investigated male attitudes towards vasectomy. Carey and Smith[316] investigated psychosocial factors affecting people infected by HIV, using the military as a target population, and Powell and colleagues[317] asked both users and providers of mental health services about service provision. Wilkinson[318] investigated interaction between women suffering from breast cancer and Stevens[319] looked at experiences of lesbian women in health. Both Keller and colleagues[320] and Smith and West[321] explored the healthcare needs of the elderly. Attitudes towards England's health strategy were investigated by Bradley and colleagues;[322] Dolan and colleagues[323] looked at how much people's views changed after deliberation; and Ramirez and Shepperd[324] explored attitudes towards different risk factors associated with cancer. The final example is provided by Weinberger and colleagues[325] who were concerned with researcher consistencies when rating transcripts.

Concept mapping

This method has been suggested by Southern and colleagues[326] as an alternative to using focus groups. It is very similar but ensures that all members have an equal opportunity to express opinion by eliminating the problems that group dynamics can cause. Additionally, it uses a combination of qualitative and quantitative data collection methods. A series of maps are developed to establish links between ideas and suggestions expressed. There is feedback provided to the participants during the process. This provides an opportunity for respondent validation, therefore increasing the overall validity of the process. Trochim and Linton[327] suggest that this is an appropriate framework for decision-making, and therefore may be particularly useful for setting priorities.

Citizens' juries

Citizens' juries consist of a group of people, representing the lay population, which discusses issues on the basis of evidence provided by experts. Citizens' juries usually occur over a period of 4 days and consist of a small group of 12–16[328] or 10–20 members.[329] Participants are selected using random and stratified sampling to be as representative of their community as possible. They should include one[329] or two trained moderators.[328] They involve members of the public in decision-making, rather than simply asking their opinion, thus taking the process a step further. The jurors are briefed about the topic in question, are given written information and listen to evidence from witnesses. Information presented to the jurors must come from several points of view and be presented in a fair way.[330] This is normally achieved by using witnesses to illustrate their cases. The members are then given the opportunity to cross-examine the witnesses and are able to discuss aspects of the subject in smaller groups. A moderator should be present for these discussions, but may not be present for the final decisions;[330] the jurors' verdict does not need to be unanimous and is not binding. A report is produced which describes the deliberations and conclusions/recommendations of the jurors and is submitted (after the jurors have had the chance to examine it). Five pilot studies were conducted in 1996,[328] and during this series the method was evaluated for effectiveness of gathering the public's opinions on priority setting[328,331,332] and specific health issues were debated, namely mental health and palliative care issues as well as payment methods for health services.[331] Additionally, three juries were held by The King's Fund, which dealt with issues concerning gynaecological cancer, back pain and who should perform certain treatments.[333] Also, the Welsh Institute for Health and Social Care (WIHSC) used this technique to explore new genetic technologies;[334] juries were also held in Portsmouth and Nottingham[328] and Lewisham.[335]

Consensus panels

These are a simplified version of citizens' juries, and consist of small groups brought together to discuss a particular issue. These groups are provided with limited information on specific scenarios. The participants are then asked to give reasoning behind the choices that they have made. Lengthy discussions then take place and are led by one of the panel members, and observed by a researcher. This technique has been applied to many different health themes, including the opinions of patients, the general public, general practitioners (GPs), specialists

and health insurers to cuts in healthcare expenditure,[336] birthing centre provision,[337] reasons for the under-treatment of depression,[338] synthesis of best practice for health technology assessment[339] and treatment for lower back pain.[340]

Public meetings

Public meetings are a common way to gain public opinion relating to health service issues. Plans and proposals are sent to CHCs for comment, and meetings are arranged to obtain a feel for attitudes towards these.[329] Many health authorities use public meetings as a means to provide information to communities and to gauge public opinion on health issues. Gundry and Heberlein[341] investigated the types of people that attend public meetings in order to test their representativeness and Broadbent[342] investigated those issues surrounding advertisement and access to public meetings in her local area. Finally, Gott and Warren[343] investigated a neighbourhood health forum that was interested in increasing local participation in health issues.

Nominal group technique

A consensus method that is similar to the Delphi technique is the nominal group technique. This was developed to avoid the problems of group interaction and provides a quantitative measure of qualitative data. In one meeting ideas can be generated and problems solved.[344] It takes the form of structured meetings facilitated by a third party and major issues affecting a particular group are identified and ranked. This method is stated as being appropriate for prioritising interventions and effective at obtaining consensus.[345] There are a number of steps in this highly structured process. Participants in the group write down their own ideas, the ideas of each participant are then listed one by one, and these suggestions are discussed and grouped into clusters. Each participant then ranks the ideas in order of importance, the ranking is discussed and ideas are re-ranked as a group. Feedback of the final results is provided.[346–348] Redman and colleagues[349] investigated use of the nominal group technique to look at priorities in breast cancer service provision in Australia. A range of health professionals, patients and relatives of patients were invited to attend the workshops: 274 agreed to participate. Care was taken to include women who were non-English speaking, and they made an important contribution to the study.

Discussion and conclusions

A number of quantitative and qualitative techniques have been identified. Whilst some are well established in terms of their use in healthcare, other new approaches have also been identified. Some of these seem to be particularly applicable for measuring public preferences.

A number of quantitative techniques were identified which have been widely used in healthcare. These include simple ranking exercises, rating exercises, satisfaction surveys and methods for estimating quality weights within the QALY paradigm (visual analogue, SG and TTO). CA (ranking, rating and discrete choices) and WTP are being developed within the context of healthcare. Techniques with limited applications in healthcare include the QDP, MoV, AHP and allocation of points.

Of the qualitative techniques, one-to-one interviews and focus groups have been widely used in many different fields and more recently in healthcare, with a current focus on their uses in priority setting. Other methods such as citizens' juries are relatively new, having been recently developed in the context of decision-making. In addition, whilst case study analysis, concept mapping and nominal group techniques are in the early stages of development, they present unique ways of analysing data collected in a qualitative format.

In the next chapter, the criteria by which techniques should be addressed is discussed, and in chapters 5 and 6 the techniques identified are reviewed according to these predefined criteria.

Chapter 4

Criteria for assessing the methodological status of techniques

In this chapter we attempt to summarise the criteria that has been used to evaluate the quality of quantitative and qualitative research. This was done by identifying both the methodological issues addressed in the papers identified in the systematic literature review (chapter 3) and current debates in the literature concerning appropriate criteria with which to judge techniques. From this the project group decided on a set of criteria for evaluating the techniques identified in chapter 3. Given the different nature of the techniques used, whilst an attempt was made to evaluate quantitative and qualitative by the same criteria, the final criteria were slightly different.

Criteria for assessing quantitative techniques

Key references used, in addition to articles identified in the review, were Brazier and colleagues[253] from their recent review of health status measures in economic evaluation, and Streiner and Norman[350] from their book on developing and using health measurement scales. The quantitative techniques identified in the systematic review of methods for eliciting public preferences, together with these two references, identified a number of criteria that have been used to assess the value of techniques. It should be noted here that different disciplines used different definitions of criteria, and we have attempted to provide an overview. Some readers may therefore disagree with our definitions. However, we hope that we have covered the main criteria.

Validity

A common criterion mentioned in a number of studies was that of the validity of the technique used. A valid measurement is one that measures what it is supposed to measure.[351] Three types of validity were identified: content validity, criterion validity and construct validity.[352]

- **Content validity** refers to the extent to which a measure takes account of all things deemed important in the construct's domain.
- **Criterion validity** (or external validity) is concerned with whether the measure adopted measures what the researcher is trying to measure.
- **Construct validity** – there are two types: **convergent validity**, which measures the extent to which the results are consistent with other measures that are held to measure the same construct; and **theoretical validity**, which assesses the extent to which the results are consistent with *a priori* expectations.

Reference has also been made in the literature to factors such as framing effects[353–356] and strategic biases,[253,257] the presence of which may cast doubt on the overall validity of the technique.

Reproducibility*

This is defined as repeatability of results over a given time. This is usually assessed as test–retest reliability whereby a sample of respondents repeat the same exercise after a short time period and their results are compared with those first time around. For continuous variables this is measured using a correlation coefficient (r),[46,253] whilst for discrete responses the Kappa coefficient (κ) is used.[357] It should be noted that implicit in this measure of reliability is an assumption that preferences exist (are complete) and are stable over time. If these assumptions are violated (which may well be the case in healthcare, where individuals are not used to making choices), then poor reliability may be indicative not of a poor technique but rather of the fact that respondents do not have complete or stable preferences. This is discussed in more detail in chapter 8.

Internal consistency

A number of studies were identified which had addressed the issue of internal consistency. This refers to the extent to which respondents have answered questions in a logical manner. The

* Inconsistency and reproducibility are consistent with Fitzpatrick and colleagues'[388] definition of reliability.

articles identified in the literature review indicated that this has been tested in a number of ways in the literature:

- One common test for consistency of responses is to include **dominant options**, i.e. if two healthcare interventions/services are being valued, and one is obviously superior to another, then the superior one should have a higher weight/score/rank.[61,196,197,199,203,205,208,253,358–365]
- Tests were identified in the literature where preferences elicited using one technique were then compared to those elicited using a second technique to see if they gave consistent results.[262,355,366–369]
- Some techniques test for consistency with respect to the stability of preferences within a task. Here individuals are presented with the same choices/items/tasks within a survey. A **judgement reliability coefficient** measures the extent to which responses are correlated.[67–70]
- Alternatively, attitude-type surveys often include statements intended to measure the same construct, and then tests are conducted to see if responses move in the same direction. This measure is usually expressed as **Cronbach's alpha** (α),[80,87–91,167,173,176–178,180,182,370] the **coefficient of reproducibility**[95,98,102] and the **coefficient of scalability**.[95,102]
- One test of consistency used in surveys is that of **transitivity**. This states that if there are three programmes being valued (A, B and C), and a respondent states that they prefer A to B and B to C, then they should also prefer A to C.[193]

Acceptability to respondents

Acceptability to respondents was identified in the literature as an important criterion. The argument posed here is that the questions posed by the survey technique should be realistic, such that respondents can answer them in a way which reflects their true preferences. The acceptability of the technique may be assessed by factors such as the questions posed, time to complete, response rates and completion rates[†].

Cost

Some studies made reference to the cost of the technique. Cost is defined broadly to include the financial, administrative and analytical burden of the technique to users. The costs to the researchers (or the commissioners of research) can be split into administrative resource use and analytical resource use. The reference to this criterion reflects the fact that whilst cost does not indicate anything about methodological quality, it represents a potential barrier to use of techniques, i.e. individuals concerned with eliciting public preferences may not use a CA exercise because they do not have the necessary expertise to do this.

Criteria for assessing qualitative techniques

Whilst it is generally accepted that quantitative research can be evaluated according to a set of predefined criteria, and there appeared to be some consensus in the literature about what these criteria were, although the terminology varied across disciplines, it became clear from the qualitative literature that there are conflicting theories/approaches on the optimal way to evaluate qualitative work.[13] Four general views were identified:

All research perspectives are unique and each is valid in its own terms

This view argues that producing lists of 'universal' quality criteria is unrealistic for qualitative analysis. They are seen as being restrictive and producing sanitised results.[13,371,372] According to this view, when using quality criteria or checklists there is the risk of producing a general consensus rather than true findings. This in turn reduces the value of the research. Here it is advocated that judgement be used with each piece of work and that the method of data collection is the most important criterion. It is stated that the data yielded must be taken in the correct context and that how the data are gathered is crucial, including the role of the researcher. This view rejects the notion of establishing quality criteria and, therefore, has little support from applied health (service) researchers,[373] commissioners and research funders.

Qualitative and quantitative methods both can be assessed but not by the same criteria

The second view advocates that it is possible to assess the quality of qualitative techniques by predefined criteria, but different criteria to quantitative must be used. Lincoln and Guba[374] suggest four criteria for evaluating qualitative research:

[†] It is recognised that such factors may not be driven so much by the nature of the technique but rather by factors such as complexity of scenarios presented and question framing.

- **credibility** – which refers to some kind of endorsement of the research findings by the research participants
- **transferability** – similarities exist between different settings and it is possible to develop working hypotheses which have potential for some transfer between settings
- **consistency or dependability** – ability to track findings to their source
- **confirmability** – providing an audit trail to allow other researchers to examine the process whereby the original researchers arrived at their results‡.

Whilst these criteria have been argued to be different to those employed by quantitative researchers, they have parallels with validity, generalisability, reliability and objectivity/ neutrality, respectively.

Qualitative and quantitative methods can be assessed by the same (or very similar) criteria

The third view is that qualitative and quantitative research are different approaches to representing the same reality. Therefore it is possible to judge the quality of both sets of research methods against a common standard, although the assessment may have to be modified to take into account the practical differences between the two approaches.[13,371,375] The majority view indicates that the main criteria to ensure quality are: acceptability; cost; validity; reliability; generalisability (or relevance) and objectivity. Most of these criteria apply also to the valuation of quantitative techniques and were defined above. However, it is recognised that some of these criteria require some modification when applied to qualitative research.[371]

Acceptability

Acceptability as a criterion is important in a culture that is dominated by statistics. Acceptability of qualitative methods amongst users, policy-makers and managers will be an important factor in the decision on which method to use.

Cost

Using qualitative methods can be expensive but, despite this, cost is not generally included in the list of qualitative criteria. Choosing a method merely because it is of low cost is not acceptable. For example, if the only reason for carrying out a focus group instead of individual interviews is because it will be cheaper, then that money can be seen as wasted if it compromises the quality of the research outcome. What is esssential is that, in assessing all types of research methods, there is some estimate of the likely value for money: will a more costly method produce even more useful findings?

Validity

Validity was defined above as measuring what you intend to measure.[351] Within the qualitative literature a number of approaches have been proposed to assess validity:

Triangulation

Triangulation originally referred to the use of three or more different research methods. Its development was a response to the recognised need to guard against errors in validity by testing whether consistent findings emerge from different methods.[7,9,10,376,377] This mixing and matching of methods has also been referred to as **methodological pluralism** and **multiple research approaches**. Whilst it originally meant using three methods at the same time, it is presently used more loosely for any study using more than one technique, on the same population, addressing more or less the same issue. Two approaches have been adopted in the literature:

- **simultaneous triangulation** – refers to using a number of techniques at the same time, on the same population, addressing the same issue, and seeing if they come to the same conclusions; to this extent it can be viewed as equivalent to convergent validity within the quantitative criteria
- **sequential triangulation** – possibility of using different methods sequentially, such that each provides a basis for the development of subsequent stages of the research process.[377]

Denzin[11] has expanded on this and has identified four types of triangulation: (i) data triangulation; (ii) investigator triangulation; (iii) theory triangulation; and (iv) methodological triangulation. These take triangulation beyond merely a method of analysis and validation of results and highlight its use as a method of enriching the research throughout the entire process.[10]

Miller and Glassner (in Silverman[378]) used the metaphor of a **bridging method** to describe a way

‡ These definitions are taken from Lincoln and Guba.[374] It is recognised by the authors that the definition of consistency appears the same as that for confirmability.

of forming links between different types of analysis. She focused upon "using several methodological strategies to link aspects of different sociological perspectives, not to discover indisputable facts about a single social reality", something that she sees triangulation as doing.

It should be noted that problems identified in the quantitative literature concerning convergent validity have also been raised within the qualitative literature. For example, if a number of methods provide contradictory results, how is it decided which results are to be used?[379] Similarly, if all the data gathered using different research methods lead to the same conclusion, how can we know that these are 'right'? It is possible that another method still would produce different results.[375]

Further, triangulation is often seen as an attempt to make the research process more scientific but, in doing this, one might lose the personal perspective.[11] Murphy and colleagues reject triangulation, concluding that "a clear exposition of the process of data collection and analysis, in which the data are related to the circumstances of their production, is essential to the evaluation of findings from qualitative research".[371] Thus, the importance of the validity of each methodological process rather than the need to replicate findings by different methods appears to be the dominant expert view of the contribution of triangulation.

Respondent validation
Respondent validation involves asking participants whether they agree with the findings of the study. If they agree then support is provided for the validity of the technique. If they do not agree, new refinements may be made to existing methods. Such a method of validation could also be applied in quantitative research. This approach has been proposed as the best way to test the validity of a qualitative study.[374] However, MacPherson and Williamson point out possible pitfalls of this technique, such as the fact that respondents' views and comments may not be easily incorporated. They recommend that "respondents in any validation exercise have a responsibility to abide by the agreed 'rules' of the process, or decline to take part, stating why, and accept their position being recorded in the final report".[380]

Reflexivity[§]
Reflexivity is a process where researchers continually reflect upon how their own interests and potential biases could alter the interpretation of the results.[308,371] Researchers come from different backgrounds, cultures and have different ideas, all of which may affect the types of questions posed. It is therefore useful to the reader of qualitative studies if the investigator describes background and training so that he or she can decide how much these factors affect the piece of research. The underlying concept is that no researcher is neutral or totally objective and therefore a reader should attempt to ascertain their role in the research process.[372] Mason strongly advocates the use of this technique in every decision made by researchers in order to strengthen the trustworthiness of the qualitative method.[372] Additionally, Banister and colleagues suggested that continuous discussion with colleagues is both stimulating and challenging and aids in extending understanding and clarity.[381] Also, in order to make explicit how ideas were formed, the keeping of a detailed diary is extremely valuable. In this way the reader of the piece of work can judge for himself or herself the context in which an assumption was made and, in turn, how valid that assumption is.

Again, this type of approach may be useful in quantitative research that is carried out in an interview setting.

Data analysis
Data analysis of qualitative data is as vulnerable to accusations of variability as are its methods of data collection, and it is vital that the method of data analysis be sound. The research will be seriously flawed if the data are not analysed in a systematic and rigorous way, regardless of how carefully the data were gathered. Miller and Glassner (in Silverman[378]) state that there is always a problem with re-telling an account of an event because it is fragmented and could be taken out of context. The process of analysis, coding and categorising, for example, means that only parts of accounts are told and cannot be viewed by the reader as a whole. The meanings of what the respondent says may be further misinterpreted by the reader him/herself. This is particularly true if the reader does not belong to the same primary group as those being interviewed, where social or cultural differences are evident or where language use is unconventional. Qualitative research has been criticised for producing vast quantities of data lacking in structure.[382] There are a number of techniques that may help to guard against the problems of data analysis.

[§] Reflexivity is relevant to both validity and reliability.

A number of **computer packages** have been designed to aid the analysis of qualitative data. As qualitative research tends to produce large volumes of data, this is particularly helpful in reducing the time taken for analysis. However, the use of computer packages has been criticised for being impersonal. A computer has no 'feel' for the results and may therefore produce less interesting results.[308] There may also be a problem with loss of data due to computer error, and it may introduce problems with confidentiality as there is the potential for more people to have access to the data set.[383] Nonetheless, it is a trend which is likely to continue, and there is a need to assess critically the benefits and costs of new packages as they appear, and the comparative advantage, especially for small studies, of computerised versus manual analyses. An important point to note here is that, whilst quantitative research is usually analysed using computer packages, problems still arise since different methods of analysis may give different results (e.g. probit analysis versus logit analysis of discrete variables).

Qualitative data analysis can be done by methods other than qualitative computer software packages. One example of this is the **framework method**, which has been developed by Ritchie and colleagues (personal communication) at the National Centre for Social Research, London. This method of analysis is worth considering.

Grounded theory is a process that reveals theory from data collected.[9,384] This "is a style of doing qualitative analysis that includes a number of distinct features, such as theoretical sampling, and certain methodological guidelines, such as the making of constant comparisons and the use of a coding paradigm, to ensure conceptual development and density".[385] The emphasis is on organising in a formalised manner the ideas that are put forward by the participants. A strength of this method is that the researcher can shift the focus of the study as the data are collected. It is often the case in qualitative work that the outcome will be unknown at the start, so the direction it takes may change in accordance with the kinds of information gathered. Bowling argued that if this technique is used "the data are collected and theories and potential concepts and categories are developed during the process, more data are collected and the theories, concepts and categories tested and so on until an understanding of the phenomenon is achieved".[9] This is a widely accepted method of analysis and serves to improve the rigour of a qualitative piece of research. It may also be useful in the area of quantitative research.

Analytic induction has also been proposed within the context of qualitative research. The aim of this type of evaluation is to identify 'deviant' cases, and to move the theory forward until it can explain all or most of the subjects being studied.[10,386] This technique begins with a rough idea or definition of a problem. The data collection process begins on one or two cases and then the original definition is reassessed, taking into account what has been realised from the cases examined. More cases are studied until there is a correlation between the cases and the theory. Deviant cases must be identified and the cyclic process repeated until all cases fit the theory. This is a very demanding and time-consuming form of analysis but is recognised as adding value and rigour to qualitative research.[308] This approach can be seen to be similar to the internal validity approach in quantitative research.

Reliability

As stated above within the context of quantitative research, a reliable measurement is one which when repeated gives a similar result on each occasion.[9,10,351,387] Within the qualitative literature, reliability is often presented as a combination of reproducibility (test–retest reliability) and internal consistency/transitivity.[388] However, Kirk and Miller define three types of reliability relating to qualitative research:[389]

- **quixotic reliability** – examines how far the method can continuously lead to the same results (which may be defined as test–retest reliability)
- **diachronic reliability** – the stability of the observation over a period of time (if this period of time is within the same questionnaire, this may be defined as internal consistency as measured by the judgement correlation coefficient; if this period of time is across points in time this may be defined as test–retest reliability
- **synchronic reliability** – the consistency of results gathered at the same moment in time but by using different techniques (which may be defined as convergent validity or sequential triangulation).

Methods to improve reliability have been put forward by Kirk and Miller.[389] When data gathering, it is advisable during note-taking to standardise the method of recording and agree on, for example, a number of punctuation signs. To avoid confusion during analysis, the ideas from the observed and the observer (researcher) should be noted separately.

Generalisability

The criterion of generalisability is concerned with whether wider claims can be made on the basis of the research.[372] Generalisability theory recognises that in any research context there are infinite sources of error. In quantitative research, sampling and statistical techniques exist to correct for possible errors.[350] For this reason it did not feature as a main criterion in the quantitative literature. Such approaches may be more difficult to adopt in qualitative research. An alternative approach, more commonly employed in qualitative research, is theoretical sampling. Here the sample chosen for the study is held to be representative of the factors that may explain variations in the subject matter being studied. If a grounded theory approach is adopted then the study sample may be extended as the study proceeds. However, others argue that once an attempt is made to generalise findings, the context is lost and it is very difficult to state whether the findings would be applicable in another settings.[10] Lincoln and Guba stipulated that generalisations spanning different contexts are impossible but suggest that conclusions drawn can be comparable and transferable across different settings,[374] provided that the settings are described and understood.

Objectivity

Objectivity has been argued by some researchers to be a fundamental part of the research process,[8,9] although other qualitative researchers have disputed this, arguing that it is a highly contentious epistemological position[#]. Using qualitative methods, the process and outcomes of the research can be seriously affected by the prior knowledge, experience and opinion of the researcher. Indeed, it is impossible to achieve true objectivity; any professional will enter into a project with preconceptions of what the outcomes may be, and to remain entirely neutral and objective is an unrealistic goal.[9] There are methods that can be employed to maximise objectivity: these are reflexivity (see page 20) and intersubjectivity.

Intersubjectivity is an approach whereby one or more additional researchers are brought in to assess whether or not they achieve similar findings from the data set as does the original researcher.[386] In order to achieve intersubjectivity (which is, of course, not the same as objectivity) in the analysis of qualitative data, more than one researcher needs to code and analyse the content of a particular data set. Each must be allowed complete freedom to select the different categories for analysis. After going through the same content analysis procedure researchers can then compare notes. Discrepancies in the coding need to be checked and negotiated until an agreement is reached. Thus, the coders are developing coding schemes independent of each other. The coding categories were developed as the coding progresses, starting with the first document of the series. To some extent this still leaves open the issue of objectivity of the actual content analysis, in ways similar to the debate about triangulation. More than one researcher analysing the same data set reduces the subjective element of a study; in other words, it may lead to a higher level of objectivity.

It should be noted that such an approach might also be used in quantitative research, where interviews are being used.

Checklists of good practice (instead of formal criteria)

Finally, an alternative approach to using predefined criteria in assessing the quality of qualitative techniques is to use checklists of good practice. It is important to note that such checklists focus on how the research was conducted, rather than on the quality of the technique itself. As such, they may be seen as vehicles to ensure validity, reliability, generalisability and objectivity. Within quantitative research such checklists have not been widely used, although guidelines were identified within the context of conducting WTP studies[390–392] and carrying out economic evaluations more generally.[393] However, within the context of qualitative research such guidelines have been argued to be valuable in establishing quality as long as they are used in an "open and permissive way".[375] A number of checklists have been created for the purposes of evaluating qualitative research.[13,394,395] Three such lists are shown in *Boxes 1–3*.

The critical appraisal skills programme (CASP), a teaching resource for evidence-based healthcare, formulated ten questions for the appraisal of qualitative methods.[394] Mays and Pope[395] published a summary of the principal approaches to evaluation in general in 1995, and updated this very recently.[13] One of the differences between these two lists proposed by Mays and Pope is that the latter set is wider on the general

[#] We thank one of the anonymous referees for pointing this out to us.

BOX 1 Checklist for qualitative methods – critical appraisal skills programme (CASP)[394]

1. Was there a clear statement of the research aims?
2. Is a qualitative methodology appropriate?
3. Was the sampling strategy appropriate to address the aims?
4. Were the data collected in a way that addresses the research issue?
5. Was the data analysis sufficiently rigorous?
6. Has the relationship between researchers and participants been adequately considered?
7. Is there a clear statement of the findings?
8. Do the researchers indicate links between data presented and their own findings on what the data contain?
9. Are the findings of this study transferable to a wider population?
10. How relevant is the research?

BOX 2 Checklist for qualitative methods (Mays & Pope, 1995[395])

1. Overall, did the researcher make explicit in the account the theoretical framework and methods used at every stage of the research?
2. Was the context clearly described?
3. Was the sampling strategy clearly described and justified?
4. Was the sampling strategy theoretically comprehensive to ensure the generalisability of the conceptual analyses (diverse range of individuals and settings, for example)?
5. How was the fieldwork undertaken? Was it described in detail?
6. Could the evidence (fieldwork notes, interview transcripts, recordings, documentary analysis, etc) be inspected independently by others?; if relevant, could the process of transcription be independently inspected?
7. Were the procedures for data analysis clearly described and theoretically justified? Did they relate to the original research questions? How were themes and concepts identified from the data?
8. Was the analysis repeated by more than one researcher to ensure reliability?
9. Did the investigator make use of quantitative evidence to test qualitative conclusions where appropriate?
10. Did the investigator give evidence of seeking out observations that might have contradicted or modified the analysis?
11. Was sufficient of the original evidence presented systematically in the written account to satisfy the sceptical reader of the relation between the interpretation and the evidence (e.g. were quotations numbered and sources given)?

BOX 3 Checklist for qualitative methods (Mays & Pope, 2000[13])

1. **Worth or relevance**: was this piece of work worth doing at all? Has it contributed usefully to knowledge?
2. **Clarity of research question**: if not at the outset of the study, by the end of the research process was the research question clear? Was the researcher able to set aside his or her research preconceptions?
3. **Appropriateness of the design to the question**: would a different method have been more appropriate? For example, if a causal hypothesis was being tested, was a qualitative approach really appropriate?
4. **Context**: was the context of setting adequately described so that the reader could relate the findings to other settings?
5. **Sampling**: did the sample include the full range of possible cases or settings so that the conceptual rather than statistical generalisations could be made (i.e., more than convenience sampling)? If appropriate, were efforts made to obtain data that might contradict or modify the analysis by extending the sample (e.g. to a different type of area)?
6. **Data collection and analysis**: were the data collection and analysis procedures systematic? Was an 'audit trail' provided such that someone else could repeat each stage, including the analysis? How well did the analysis succeed in incorporating all the observations? To what extent did the analysis develop concepts and categories capable of explaining key processes or respondents' accounts or observations? Was it possible to follow the iteration between data and the explanations for the data (theory)? Did the researcher search for disconfirming cases?
7. **Reflexivity of the account**: did the researcher self-consciously assess the likely impact of the methods used on the data obtained? Were sufficient data included in the reports of the study to provide sufficient evidence for readers to assess whether analytical criteria had been met?

worth of the research (e.g. *Box 3*, point 1), whilst the first list is more methodological.

Discussion and conclusion

Following the criteria identified in the quantitative literature, the following five criteria were used to evaluate quantitative techniques:

- validity

- reproducibility
- internal consistency
- acceptability to respondents
- cost.

In addition, three additional properties of techniques were thought to be important to the research team: (i) whether the technique has a **theoretical basis**; (ii) involves a **constrained choice**; and (iii) results in a measure of **strength of preference**. These additional criteria may be seen as properties of the technique itself, and will not be influenced by the nature of the study carried out. Theoretical basis refers to the extent to which the technique has some theory behind it so that the axioms on which it is based can be tested. Constrained choice refers to the extent to which the technique takes account of the real decision-making context facing decision-makers, therefore incorporating the concepts of scarcity and opportunity cost. Strength of preference is interpreted as whether the tech-nique indicates intensity of preference between choices; in other words, whether the technique gives a cardinal rather than ordinal ranking of options[¶].

The issue of quality in qualitative research has clearly received increased attention in recent years. However, there seem to be no set guidelines, with different researchers adopting different approaches. Four approaches were identified. The first view, that such research cannot be evaluated, is unlikely to receive much support from methodologists or funders of research. The second and third views suggest that criteria can be applied, but that these criteria may be either different to those for quantitative, or the same, but operationalised in a different manner. In looking at the second and third views it is clear that many of the criteria used have some parallels with quantitative criteria. For example, simultaneous triangulation is very similar to convergent validity, and analytic induction has close ties with internal validity. Quantitative research may also gain by extending definitions of validity to include respondent validation and reflexivity, and applications of quantitative techniques may be improved by the development of checklists for good practice. Finally, checklists have also been proposed as a method for evaluating qualitative research. We adopted the third view. Based on the literature and discussion with colleagues, six criteria were decided upon to evaluate the qualitative techniques identified:

- validity
- reliability
- acceptability
- cost[**]
- generalisability
- objectivity.

Chapter 5 presents the review of the quantitative techniques identified in chapter 3 and chapter 6 the review of the qualitative techniques identified in chapter 3.

[¶] In an attempt to subject the criteria to peer review to ensure that the criteria being used to assess instruments were appropriate, a questionnaire was administered to a convenience sample of 67 researchers with some knowledge of methodology and who may be involved in eliciting public preferences. The intention here was to assess whether the initial set of criteria covered all factors that were important to users of research instruments. The sample was taken from staff at the University of Aberdeen (including members of the Departments of Public Health, General Practice, Sociology, and Education, and the Health Economics Research Unit and Health Services Research Unit), staff of Grampian Health Board (Acute Services Team and Primary Care Unit) and members of the Health Economic Study Group (at their bi-annual conference). The results indicated that the criteria chosen were all judged to be important, and no additional criteria were mentioned.

[**] It is noted here that cost is not an evaluation criteria *per se* but is an important consideration and component of whether a particular technique is used.

Chapter 5

A review of quantitative techniques for eliciting public views

In what follows, the quantitative techniques identified in chapter 3 are reviewed according to the criteria in chapter 4. The presentation is qualitative in nature. Appendix 4 presents an initial attempt to provide a quantitative synthesis by providing weights for the different criteria. This will be developed in future work.

An important distinction that should be made when evaluating quantitative techniques is that between an approach to and an instrument for eliciting preferences. This is an important distinction when considering the extent to which conclusions regarding whether techniques satisfy certain criteria are generalisable. For techniques such as SERVQUAL, SEIQoL and generic satisfaction surveys it is possible to establish the methodological status of the instrument generally within the patient group and context explored. This is because of the instruments' fixed content and design. However, for general approaches to eliciting preferences, such as CA or WTP, satisfaction of the criteria will depend to a large extent on the way the research is conducted. Given this, whilst it is possible to get some overall feel for the satisfaction of the criteria for the approaches, such conclusions are not as strong as those conclusions made regarding specific instruments*. Further, whilst the criteria of validity, reproducibility, internal consistency, acceptability to public and cost will be determined by the way the research is conducted, the criteria of theoretical basis, constrained choice and strength of preference can be seen as properties of approaches and instruments and therefore will not be influenced by the research.

In the process of the evaluation exercise it became clear that four of the techniques identified in our review (visual analogue, SG, TTO and PTO) had been extensively reviewed by Brazier and colleagues in their recent HTA monograph.[253] We obtained the references listed in the Brazier report, reviewed them in the light of our criteria, and updated the literature identified by Brazier and colleagues with studies and issues raised from our literature review which had not been identified by Brazier and colleagues.

Ranking techniques

Approaches to eliciting preferences

Simple ranking exercise

There is no well-defined theoretical basis for simple ranking exercises, and the output of such studies is clearly ordinal. Whilst the types of questions posed may be argued to be relatively simple (compared with some techniques that will be introduced later), the usefulness of the output of such questions to addressing issues of resource allocation has been challenged.[396] The main limitation is that it is not clear what decision-making rule should be implemented from the results of a ranking exercise. For example, in the British Medical Association (BMA) study, reported by Groves,[17] from a list of ten healthcare interventions, the public ranked childhood immunisation at the top of the list, followed by GP care for everyday illness, screening for breast cancer, intensive care for premature babies, heart transplant operations, support for carers of elderly people, hip replacements for elderly people, anti-smoking education for children, treatment for schizophrenia and finally cancer treatment for smokers. Therefore, should all children be immunised? And then GP care be provided for all everyday illnesses? And then we move down to the third priority – should all women be screened for breast cancer? From this example the obvious limitations of a simple ranking exercise are clear. The main problems relate to a lack of consideration of the marginal context of decision-making, the lack of consideration to the principle of opportunity cost, and the failure of the technique to provide a measure of strength of preferences.[17] Bowling and colleagues[16] highlighted the high cost of interview. Potential costs of studies were cited to range from £20,000 to £90,000. The

* We would like to thank an anonymous referee for making this point.

authors further state they did not receive the cooperation of health councils who protested to the ethics of the study (this may have happened for any of the techniques used).

Response rates have generally been good for interviews,[16,18,24–26,366] but more varied for postal questionnaires.[16,22,366] Bowling and colleagues[16] reported a response rate of only 11% for a public sample, after four mailings, and 66–68% for doctors, again after four mailings, whilst Kinnunen and colleagues[22] report postal response rates of 52–68%, including 59% for their general public sample. Bowling,[366] in a sample of 350 respondents, found item response to vary between 290 and 350.

A limited number of studies were identified that had addressed internal consistency or reproducibility. Shickle[358] found evidence of internal inconsistency in responses: whilst respondents gave their highest rankings to life-saving treatment, they stated a preference for quality of life improving treatments. Bowling[366] found identical rankings in test–retest reliability within their pilot samples, but could not compare with the main sample because of changes in wording.

A number of studies have, however, addressed various aspects of validity. Several studies have supported internal validity through confirming hypotheses that respondents favour the young over the old, females over males, non-smokers over smokers, non-drinkers over drinkers, and poor people over rich.[17,18,25,26,358] Furnham and Briggs[26] further found that older people would be more likely to discriminate over place of birth and that women would be more likely to prioritise those with dependants. Bowling,[16] in a priority ranking of 12 health services, observed that the priority ranking of both their public samples (community groups and a postal survey) were identical or very similar in the majority of cases. They further found GPs' and consultants' rankings also to be very similar to each other. However, the authors' pilot demonstrated that responses were sensitive to question wording. For example, one of the health services, "intensive care for premature babies" was given the highest priority (rank 1) but when it was qualified with the statement "weighing less than one and a half pounds and unlikely to survive" it dropped to rank 10. Concerning convergent validity, Angermeyer and colleagues[24] found a ranking exercise led to more prominent differences in recommendations concerning schizophrenia than a Likert scale.

Qualitative discriminant process

The QDP is in the early stages of development, there have been no applications eliciting public views in healthcare, and little methodological work has been carried out. However, within the context of healthcare, the technique does have appeal. Bryson and colleagues[28] note that one of the significant advances of the technique over other methods of eliciting preferences is its ability to deal with the vagueness that exists in human decision-making. This may be particularly attractive in healthcare, where preferences may be both vague and difficult to articulate. The procedures involved, whereby respondents are encouraged to re-think their answers, move options around at various stages and initially define preferences in qualitative terms, is attractive. However, it should be noted that the technique relies on computer-based interviews, and will therefore be relatively expensive to carry out. Bryson and colleagues[28] argue that QDP is "simple and intuitively appealing". It has its theoretical basis decision theory, fuzzy set theory, the theory of vague real numbers and voting theory. Given the mapping of qualitative responses onto an interval scale, it may be argued to posses strength of preference properties. Ngwenyama and Bryson[30] note that given the computer-based collection of data, inconsistent answers are not permitted.

CA ranking exercises

In terms of theoretical basis, the CA ranking approach was developed in mathematical psychology from conjoint measurement theory.[32] This theory argues that one of the great strengths of CA ranking exercises is that cardinal data can be obtained from ordinal rankings.[32,398] As Green and colleagues[398] note:

> "Recent developments in mathematical psychology, culminating in a new approach to measurements – called conjoint measurement – allow the researcher to measure the separate or parts-worth of each benefit to overall profile value. These approaches use various algorithms to transform the dependent variable, the preference ordering, into a cardinal (interval) scale and, in the process, develop cardinal scales for the contributory benefits that reflect their relative importance."

In addition, CA ranking exercises are rooted in Lancaster's theory of value.[31]

The ranking conjoint approach has undergone relatively little empirical work in healthcare compared with the choice-based approach. Response rates have been good with 59% reported in a

postal survey,[45] 81% in a survey of physicians (including a $50 incentive),[37] and 84% and 90% in interviews.[42,46] One study reports that respondents had little difficulty in completing the exercise,[46] two studies reported respondents ranking 18 profiles,[37,43] whilst respondents in another study took only 12–18 minutes to rank 25 profiles.[42] Whereas one study reported between 20 minutes to 1 hour to rank 12 profiles,[39] another limited profiles to a "manageable" ten.[44] No comparisons have been carried out between the CA methods, but one study argues that ranking is easier to complete than choice-based CA (see below for a description of this technique).[39] Generally, high levels of consistency have been reported,[21,41,43] but further research is needed. In the only study identified that assessed reliability, test–retest coefficients of 0.7 and 0.66 were estimated for ten respondents after a 7-month interval between questionnaires.[46] Evidence supporting internal validity has been found in terms of coefficients on criteria behaving in line with *a priori* expectations.[21,43,44,46] Further, Singh and colleagues[46] confirmed their four hypotheses concerning the influence on different groups (the risk conscious group were most influenced by side-effects, the child-focused group were most influenced by the child's attitude, the cost-conscious group were most influenced by out-of-pocket costs, and the ease-of-use group were most influenced by route of treatment) and the presence of lower income and education in the cost-conscious group.

Rating techniques

Approaches to eliciting preferences

Rating scales within the QALY literature

Supporters of the technique have argued that its attraction lies in the fact that it has a theoretical basis (in psychometrics) and is relatively easy to complete (compared with SG and TTO; see below).[248,399–402] There is, however, debate about the extent to which the VAS incorporates a measure of strength of preference.[246,399,403–405] It has been argued that given that respondents are instructed to place health states on a line such that the intervals between the points reflect the differences they perceive between the health states, the output from this technique can be regarded as cardinal.[352] However, others challenge this view.[246,399,403,404] For example, Nord states that "subjects are clearly ranking states when placing them on the scale, but they express little depth of intention beyond this ordinal level of measurement".[399]

Brazier and colleagues[253] found the VAS to be the most acceptable of the health-state valuation techniques, eliciting high response and completion rates.[352,359,400,406–411] These findings were supported by the additional literature identified in our review. Revicki[412] and Badia and colleagues[365] reported completion rates of 98.7% and 98%, respectively, whilst Morss and colleagues[413] reported a 100% completion rate.

Brazier and colleagues[253] also noted that the VAS method is less expensive to carry out (since it can be conducted via questionnaire-based surveys) and quicker to complete, but not necessarily easier to complete, than other health-state utility measures.[400,414,415] Some support was found for this conclusion in our review. Shiell and colleagues[416] observed that many patients found it difficult to rate death on the VAS. The authors' explanation for this was that the patients completed the exercise with reference only to the best imaginable health-state (full health).

The VAS has tested for internal consistency by including health states that should logically be preferred to others, and then testing whether such states are given a higher utility weight. Brazier and colleagues conclude that the VAS performs well according to this definition of consistency.[253] Similar results were found in the additional studies we identified. For example, Krabbe and colleagues[417] tested the internal consistency of five health-state valuation techniques using the above method. The average inconsistency for the VAS method was lower (2.0%) than all the other methods (closest were SG and TTO at 4.6% and 4.3%, respectively). Also, Badia and colleagues[365] show encouraging internal consistency results, citing a significantly lower percentage of logical inconsistencies than the TTO method.

Test–retest results are good for the VAS,[359,400,409,418–421] with the only exception being Rutten-van Molken.[422] These results are a hybrid of the reviews by Froberg and Kane[352] and Brazier and colleagues[253] since no additional literature was identified in our review.

With respect to convergent validity, VAS utility weights have been found to be consistently lower than those estimated using either SG or TTO.[400,412,413,422–424] Other evidence concerning the validity of the VAS is highlighted in the Brazier and colleagues report.[253] They noted that VAS methods are susceptible to response

spreading[425] whereby "the respondent seeks to place (spread) responses across the whole (or a specific) of the available scale". This is also cited in Kaplan and colleagues[407] and Revicki.[412] Brazier and colleagues[253] suggest that such findings indicate the lack of an interval scale within the VAS. The validity of the VAS has also been challenged on the basis that preferences are elicited under certainty, and therefore it is value that is being elicited, not utility. However, Dyer and Sarin[426] suggest a function that provides a link between value and utility.

Rating scales within CA studies

The CA rating approach has its theoretical basis in information integration theory and judgement analysis[427,428] where again it has been argued that cardinal data can be obtained from individual responses to rating data. In addition, CA rating exercises are rooted in Lancaster's theory of value.[31]

Postal response rates have ranged from 42% to 67%[52–54,57,61–63,66] and 33% was reported in one interview situation.[58] Completion rates have been high (78–85%).[57–59,66] Harwood and colleagues[58] noted that their questionnaires "were probably about as difficult as it is reasonable to undertake", although their response rate was comparable with other scaling methods. Graf and colleagues[56] observed that the method could result in fatigue. They noted that rating 16–24 profiles takes around 10–20 minutes and that telephone interviews using CA are difficult and the technique is best suited for face-to-face interviews. However, they note that mail surveys are possible. A high level of consistency has been shown, but in limited studies.[61,63,64] Similarly, good reliability has been found in two studies but with low numbers.[58,64] Some evidence has been provided to support internal validity[61,63] and no evidence of ordering effects has been detected.[61]

Likert scale

There is no constrained choice and debate exists as to whether strength of preference is measured. Given that there is no assumption of equal intervals on the scale, the difference between 'agree' and 'strongly agree' may be perceived by the respondent to be greater than that between 'agree' and 'undecided'.[9] Likert scales are relatively easy to complete, with response rates of 84–91% reported in interview or telephone surveys[76,81] and a wider range of 49–96% in postal studies.[73,74,77,82] Completion rates of 95% and 70% were noted by Pfennings and colleagues[83] and Marks and colleagues,[80] respectively. Marks and colleagues[80] further stated that their 20-item questionnaire could be completed in less than 5 minutes. Likert scales are often used to make comparisons between individual statements and compared across the different respondents and the different issues; however, some researchers go further and total up scores across different scales.

Methodological evaluations of Likert scales have been limited in healthcare, outside of their use in satisfaction studies. Marks and colleagues[80] found good internal consistency and reliability. Construct validity was also supported[80] and, in terms of convergent validity, Likert scales demonstrated (often significantly) higher results than VAS.[83] A major disadvantage of totalling up scores is that "while a set of respondents will always add up to the same score, the same total may arise from many different combinations of responses, which lead to a loss of information about the components of the scale score".[9]

Semantic differential technique

In a short review of the SDT, Bowles[89] suggests that the SDT is acceptable because it requires a short amount of time to complete and can be used with a variety of populations. This conclusion is supported by the work of Niedz.[140] However, Appels and colleagues[429] reported a large number of incomplete responses in two large cohort studies (59% in one cohort and 32% in the other). Like Likert scales, analysis is straightforward, requiring only comparisons of mean and standard deviation calculations. Mixed results have been found regarding the reliability and validity of the technique.[85,89,92,140,430] Girón and Gomez-Beneyto[93] challenge the techniques' validity. In a study of the relationship between family attitudes and relapse in schizophrenia, they argue that it is possible that the "semantic differential does not measure the relative's attitude but their accuracy in assessing the severity of the illness and behavioural disturbance of patients who are more likely to relapse". Swain and McNamara[92] note that the principle disadvantage of the SDT is the ease with which investigators can choose adjectives that are inappropriate to the concept domain of interest, despite the existence of methods to aid appropriate choice. Holmes[85] found that SDT suffered from two more general questionnaire effects. First, he noted a through questionnaire bias, whereby attitudes are modified somewhat from initially more extreme positions; secondly, a within page bias, whereby respondents prefer the left-hand side of the page. These limitations may be present with other techniques.

Guttman scales

The Guttman scale derives from psychology, and is rooted in 'facet theory'.[98] It neither incorporates strength of preference nor a constrained choice. Because of the lack of published studies, little can be gleaned on the methodological status of this method. Response rates of 71% and 65% have been reported, which are good for postal questionnaires,[100,102] although poorer response rates have been reported by psychologists.[95] Internal consistency is nearly always reported because of the selection of items for the Guttman scale. The reported internal consistency is no more than moderate.[99,101,102] In addition, it seems to be common practice to report a 'scalability co-efficient', which accounts for the artificial effect resulting from selecting the items for the scale. Again, this was reported as moderate, implying that the published studies were not very good at picking the right items for the Guttman scale. Reproducibility was not reported and Petersen[102] notes that a valid scale was not obtained in his study.

Instruments for eliciting preferences

Satisfaction surveys[†,‡]

Satisfaction surveys appear to lack a clear conceptual or theoretical basis, resulting in some confusion over what it is they are measuring.[143,145,384,431–435] The 'common sense model', which assumes that patients derive their level of satisfaction by comparing their experiences with their expectations, derives from multiple discrepancy theory.[145] However, this theory has been challenged within the context of satisfaction surveys.[111,145] As Carr-Hill notes: "Measured achievements-aspiration gaps … may well be rationalisations rather than causes of satisfaction ratings."[145]

Constrained choice is not normally an aspect of satisfaction studies. A further limitation of satisfaction surveys is that they provide no guidance on where scarce resources should be concentrated.[143,145] This is a result of such surveys failing to incorporate any notion of constrained choice or strength of preference. For example, suppose respondents were asked to rate their satisfaction with five aspects of care (waiting time for appointment, time in waiting room, time in consultation, information received and location of appointment) on a rating scale from 0 to 7, where 0 represents completely dissatisfied and 7 completely satisfied. Suppose the results indicated that patients were most satisfied with time in consultation, followed by time to appointment and least satisfied with time in waiting room. What then are the policy implications? Should resources be devoted to time in the waiting room since patients are least satisfied with this aspect of care? The problem with this decision-making rule is that the characteristic of care that patients are least satisfied with might also be one that they are least concerned with. Further, despite the fact that patients were most satisfied with time in consultation, they may still prefer marginal improvements in this characteristic of care over others that they are less satisfied with. To make decisions regarding the optimal allocation of resources, information is required on the weights attached to the various dimensions that make up 'satisfaction'.

Despite these properties, satisfaction surveys are clearly a popular method for eliciting preferences. This popularity may partly reflect the fact that researchers find such studies relatively easy to conduct, respondents find them relatively easy to answer, and analysts find them relatively easy to analyse. Response rates for satisfaction surveys have ranged from 38% to 91% with postal questionnaires.[73,103,109,111,113,116,123,131,142,154,156]

Whilst it is generally agreed that satisfaction studies provide a relatively low cost research method, Bisset and Chesson suggest that they "may not represent value for money"[159] and McKinley and colleagues highlighted that a well-designed and piloted satisfaction questionnaire demands "time and expertise"[436] which could make a survey expensive.

Despite their acceptability to researchers and respondents, a number of problems have been identified in the literature. Satisfaction studies target patients (service users) and their family/carers, but not the general public. The counter-argument is, of course, that those people with

[†] The NHS R&D HTA programme has commissioned a separate systematic review on patient satisfaction. This review is entitled 'The measurement of patient satisfaction: implications for health service delivery through a systematic review of the conceptual, methodological and empirical literature' (project number 96/27/02) by Crow and colleagues. This report will provide more detailed information of the methodological literature.

[‡] Satisfaction surveys may be seen as both an **approach** to, and **instrument** for, eliciting preferences, depending on whether the researcher devises his or her own instruments or uses an off-the-peg questionnaire.

experience of a particular (health) service are in the best position to comment up its quality and quantity. In addition, satisfaction questionnaires are said to "encourage patients to respond to his or her own health care on an individual basis without reference to the wider collective of healthcare users".[431] As with all questionnaire based surveys, the wording and presentation of questions in a patient satisfaction survey may influence the responses.[437]

Perhaps the greatest challenge to the validity of satisfaction surveys is the finding that patients generally report high levels of satisfaction with the healthcare they receive.[111,145,438,439] Fitzpatrick[144] notes that high levels of satisfaction are typically recorded by at least 80% of respondents to satisfaction surveys. Such high satisfaction casts doubt on the ability of such surveys to detect real differences in patients' opinions.[432] Possible reasons for this include the reluctance of patients to express negative views about their healthcare[440] – also referred to as 'gratitude bias' – and a general feeling that 'what is, must be best'.[439] Cleary and McNeill concluded in their review of the patient satisfaction literature that "... frequently only global measures of satisfaction are used".[441] However, OE and smaller scale satisfaction studies are more likely to report areas of criticism than large-scale (postal) questionnaire surveys.[437]

Avis argues that despite all the criticism of satisfaction studies, "there is a central place for satisfaction surveys in monitoring standards of quality in health care. However, they must be sensitively performed."[431] Others have highlighted the need to use research tools that have been shown to be reliable and valid.[437] There are now several 'off-the-peg' satisfaction questionnaires in use.[431] Researchers using one of these instruments can be assured that the instrument has a certain level of validity and reproducibility.

Schedule for the evaluation of individual quality of life

Only a handful of SEIQoL studies were identified, and most of those have been for small groups. Thus, little can be concluded in terms of its methodological status. Browne and colleagues[69] found a response rate of 90% in an elderly group suffering from dementia (n = 56; mean age 74 years), and concluded that these respondents were able to understand and complete the questionnaire. However, Coen and colleagues[67] found that only 6/20 patients were able to do so. Some evidence has been found for internal consistency,[67,69,70] reproducibility,[68] and construct validity[67,68] and internal validity.[69] However, once again these are with small numbers of respondents.

SERVQUAL

Although SERQUAL has received much empirical attention in the healthcare literature,[166–182] little methodological work has been carried out.[370,442–445] The original method employed 22 statements divided into five dimensions on a seven-point Likert scale. It has been observed that, although the technique is "reportedly business neutral", item statements require minor changes in wording to make them appropriate in the healthcare setting.[179,180] Further, researchers have replaced the seven-point scale with a five-point Likert scale, arguing that this leads to less frustration on behalf of respondents, and encourages higher response rates and improved quality of responses.[177,370] These authors also removed negatively worded statements which reportedly led to confusion and irritation of respondents in favour of positively worded statements only. Others have used a nine-point scale.[169,171,446] Later versions of SERVQUAL have incorporated a budget pie question (see below for description of this technique) in which respondents divide 100 points between the five dimensions[167,172,447] whilst others have used a 100-point rating scale for each item.[171] Hart[447] states that the seven-point Likert scale may not have equal intervals and respondents may

> "actually deploy an 'increasing resistance' model in which it is easier (in psychometric terms) to move from the central point (point 4) to its immediate neighbour (point 5) than it is to move from a position of near perfect satisfaction (point 6) to perfect satisfaction (point 7)".

Instead, the author suggests assigning cardinal values to each point as follows: point on scale 1 (value –6); 2 (–3); 3 (–1); 4 (0); 5 (1); 6 (3); 7 (6).

Response rates have been variable, ranging from 22% to 72% for postal questionnaires,[170,171,174,179,181,370,442] to 73% by telephone[176] and 36–80% when distributed by hand.[171,175,180] Raspollini and colleagues[181] observed that whilst it is not necessary to train interviewers, interpretation and analysis of results are "difficult" and "complex".

Evidence of internal consistency and validity were supported in the marketing literature,[165] and Dyck[448] argues that these are preserved if

"the intent and order of the questions remains the same". There has generally been good evidence of the internal consistency of the dimensions of SERVQUAL.[167,168,173,176–178,180,370] Some studies support its validity and applicability to healthcare,[167,173] whilst others dispute this, challenging its construct validity and claiming intercorrelation of its dimensions.[179,443] One study by Oswald and colleagues[442] found that dimensions of quality assessment differed from those assessed in SERVQUAL. Duffy and Ketchand[182] found a large unexplained variation in service satisfaction, concluding that dimensions that may have influenced satisfaction were excluded.

Choice-based techniques

Approaches to eliciting preferences

Simple choice exercises

In terms of acceptability, Charny and colleagues[183] reported that whilst 31% of respondents had made all 13 choices, nearly 69% were unable to complete all the choices. However, this finding may relate to the sensitive nature of the question being asked. Internal validity was also found in the related studies as respondents favoured the young over the old, women over men, non-smokers and non-drinkers over smokers and drinkers.[183,184] In the study by Charny and colleagues,[183] the instance where an 8-year-old child was favoured over a 2-year-old was explained by the greater parental investment in terms of effort and emotion.

Regarding RPS, Ryynanen and colleagues[186] observed that members of the public completed the questionnaires quickly. However, most had difficulty in making choices between two alternatives, some found that the questionnaire made them anxious and one reacted aggressively. However, these findings may reflect the sensitive nature of the questions being asked. Medical and nursing undergraduates completed the questions quickly and without difficulty. No information is available on internal consistency. Eight respondents answering three different sets of RPS questions measured test–retest reliability. The authors concluded that reliability was "good".[186] Ryynanen and colleagues[186] found comparable results in terms of criterion validity between conventional and RPS questionnaires in an undergraduate student population.

CA choice-based questions

The increased number of applications of choice-based CA has been accompanied by an investigation of many of the main methodological issues that are important when developing a technique in a new setting. Choice-based CA exercises are held to be acceptable to individuals on the basis that they present them with the types of decisions they face on a daily basis. It is this argument that has led to the choice-based technique being preferred over ranking and rating approaches.[52,60,61,194,206,449] Choice-based techniques have also been favoured by economists because of their underpinnings in a branch of economic theory known as random utility theory.[450,451] However, it has been noted in the literature that the number of choices presented to individuals should not exceed 12 (at maximum), and the number of criteria should not exceed five or six. More choices, or criteria within the choices, will result in difficulties in completing the questionnaires.

Choice-based CA studies have mainly been carried out using postal questionnaires, producing response rates ranging from 18% to 88%. Response rates are generally higher if questionnaires are sent to individuals involved in a randomised trial, include financial incentives, are accompanied by a physician's covering letter, or are given out in clinics.[188,191,193,196,197,199–203] Few difficulties have been reported when answering choice-based CA questions[191,194,199,201–203,396] and two studies report that the technique was well received by policy-makers.[198,205]

Many choice-based conjoint studies have included tests of internal consistency by including dominant options, i.e. options where one scenario is considered 'better' than the one it is being compared with on all criteria. Respondents would therefore be expected to choose this scenario. Tests of this nature suggest a low proportion of inconsistent responses.[193,194,196,197,199,202,203,205–207,396] One study was identified which tested internal consistency with reference to transitivity, i.e. if A is preferred to B and B is preferred to C then A should be preferred to C.[193] In this study, 6% of respondents failed at least one of the consistency tests.

Only one study was identified in the healthcare literature which had assessed the test–retest reliability of choice-based CA. In a study looking at parent preferences for out of hours care, San Miguel and colleagues[357] found a high level of reproducibility within a 2-month time period.

The validity of choice-based CA has been addressed at a number of levels. The method

has demonstrated high levels of internal validity.[187,190,194–197,199–203,207] Convergent validity has been with SG and VAS,[192] and WTP.[194,199] A number of studies have tested for framing effects. Whilst Vick and Scott[203] found some evidence of an ordering effect with regard to the criteria,[202] two studies have found no evidence of framing effects.[61,189] Ryan and Wordsworth[200] found WTP estimates derived from choice-based CA to be insensitive to the level of criteria.

CA assumes that individuals have **continuous preferences** such that there is always some improvement in one attribute that can compensate an individual for a deterioration in the level of another attribute. Tests are often carried out to see if this assumption holds.[193,195,199] There is less agreement in the literature concerning how to deal with individuals exhibiting non-compensatory behaviour. Some studies drop such respondents, arguing that marginal rates of substitution (the rate at which they trade between criteria) cannot be estimated for them (since they are not willing to trade) and therefore the utility function estimated is inappropriate.[194] Others compare estimated utility functions with and without non-traders.[187]

Analytic hierarchy process

The AHP has been shown to have an axiomatic foundation and to produce output on a ratio scale.[218,219,223,225] The technique typically collects data using interactive techniques (often computer-based).[452] Whilst this may increase the cost of data collection, it can also ensure the consistency of responses (see below).

Studies employing the technique, mainly in a face-to-face type of interview, suggest a high level of acceptability. Schwartz and Oren[225] reported that 83% of patients completed and understood a questionnaire that consisted of 28 pairwise comparisons, and Peralta-Carcelen and colleagues[452] reported that 90% (83/92) of pregnant women, 51% (40/78) of obstetricians and 67% (40/60) of paediatricians participated in their study. Within this study, 89% of the women and 90% of the physicians were reported to have understood the interview format, although there were some difficulties reported. However, these appear to concern the nature of the data presented to them and not the method. The authors further note the limited educational background of their sample of women but noted they did not have undue difficulty in understanding the model and thus claim that this justifies the generalisability of the method. Dolan[235] directly addressed the question of whether patients were capable and willing to use AHP in medical decision-making. Following a structured interview, respondents ere followed up with five evaluative questions concerning the technique. Despite the small numbers (n = 20), positive findings were found. In all, 90% (18/20) of respondents were defined as "capable and willing" to complete the exercise, and Dolan concluded that AHP was "likely to be acceptable to and within the capabilities of many patients". Positive findings have been found elsewhere, albeit with small numbers, as have further cited studies (published in abstract form only).[232,453–455]

Javalgi and colleagues[239] conducted a postal questionnaire to assess factors important in choosing a hospital. They justify a mail survey on the basis of cost and the nature of the data required. A 47% response rate was reported (235/500) with 220 responses being usable.

The introduction of the software has facilitated the application of the AHP in terms of the collection and analysis of data (with weights being automatically estimated).[226,235,237,239] This computer software also has in-built tests of consistency. Consistency in the AHP refers to the axiom of transitivity, i.e. if A is twice as preferable as B, and B is three times more preferable than C, then, to be consistent, A must be six times more preferable than C. A consistency ratio is produced, with a value of 0.1 or less being defined as acceptable.[227,240,452] High levels of consistency have generally been reported. Peralta-Carcelen and colleagues[452] found that of 83 pregnant women, 40 obstetricians and 40 paediatricians, only three of the women were inconsistent. Schwartz and Oren[225] excluded 5% of respondents from the analysis since their consistency ratios exceeded 0.50 (the remaining ranged from 0.11 to 0.23). Finally, Hannan and colleagues[241] argue that their responses could not have been randomly allocated.

Very little work has addressed issues related to the validity of the technique. Schwartz and Oren,[225] in a study looking at patient preferences for treatment of myocardial infarction, note that the choice model accurately predicted behaviour. Dolan[235] notes that the greatest challenge to the validity of the technique comes through the **ranked reversal phenomenon**. Here a change in the relative desirability of a criterion/alternative is caused by the introduction of another criterion/alternative. However, a new method of AHP has been developed which precludes rank reversal.[229,235]

Dolan[226] notes that this method has successfully been used to promote a change in behaviour in patient care[§].

Standard gamble

The SG technique has often been argued to be the gold standard by economists since it is rooted in expected utility theory (EUT).[49,245,352,456,457] From this theory, which explains individual decision-making under uncertainty, it is then held that the utility weights derived from the exercise represent a strength of preference measure.[458] However, given the known problems individuals have with probabilities, and the nature of the questions posed (with the individual required to specify a probability level of indifference), SG exercises must be carried out via an interview, with props, and the interviewer must be trained.[253] Given this, the technique is likely to be relatively expensive.[253,352,400]

The review by Brazier and colleagues[253] found high levels of response rates and completion rates, ranging between 80% and 100%.[253,361,411–413,459–462] Similar conclusions were derived from the additional literature we identified in our review.[365,458] Shackley and Cairns[458] reported a 95% response rate and Badia and colleagues[365] a 98% rate. It must be remembered that such rates are for interviews where completion rates would be expected to be relatively high.

Brazier and colleague[253] note that where completion problems are reported when using the SG method, these difficulties are no worse than those faced by other techniques. They highlight the problems encountered by Patrick and colleagues[463] and van der Donk[414] where it was stated that the SG was no more burdensome than either the TTO or VAS methods. However, in a study identified in our review, Hall and colleagues,[464] in comparing VAS, TTO and SG, found SG to be the most difficult technique to understand. Richardson[403] observed that "empirical evidence suggests that people have difficulty understanding the objective meaning of [SG] probabilities, especially extreme values". This is a view shared by Johannesson and colleagues[404] and Froberg and Kane,[352] who both state that some decision theorists see the SG as being too difficult to explain to respondents.

Tests of internal consistency have been carried out in the same way as for the VAS, and the results are mixed.[253,352,361,465] Dolan and colleagues[362] in testing SG against 12 logically consistent comparisons, argued that the technique produced high levels of consistency. This study incorporated a test between two variants, one involving a prop to aid the decision-maker (a specially designed board and cards) and the second involving a self-completion booklet with no props. Consistency rates for SG with props was 83.8% and 87.5% with no props. Llewellyn-Thomas and colleagues[361] also found a high level of consistency, with 84% of respondents (54/64) ranking five health states in accordance with *a priori* expectations. However, Froberg and Kane[352] argue that given that people find it difficult to work with probabilities, and the fact that they may also have an aversion to taking risks, responses are often inconsistent. This conclusion was supported by Thompson.[465]

Brazier and colleagues[253] reported good test–retest results for the SG technique.[362,400,409,418,419,422,459,466] No additional studies were identified in our review.

Mixed evidence exists concerning the technique's validity. In terms of convergent validity, Brazier and colleagues[253] reported that whilst SG has been shown to correlate reasonably well with TTO,[362,466–469] it does not correlate well with either health status or the VAS.[409,422,456] O'Brien and Viramontes[419] found that in comparing SG to VAS and WTP, SG was most highly related to the latter. An additional study we identified in our review, by Bala and colleagues,[470] found SG to be consistent with WTP.

SG is based on an assumption that individuals are willing to trade risk. Individuals who exhibit non-trading behaviour may do so either because they are completely risk averse, always preferring the certain intermediate outcome to the gamble, or because they are risk lovers, always preferring a gamble to the certain health outcome. Such behaviour would provide information concerning the value of the health state being considered. However, if non-trading is a consequence of the respondent objecting to taking risks, the values estimated from a SG experiment will not indicate utilities of given health states.[471]

[§] The question of whether to preclude rank reversal depends on the reason for the rank reversal. If rank reversal occurs because respondents do not understand the technique, this may be a reason to exclude such cases. However, if rank reversal occurs because of a change in preferences, then such exclusion is not valid. This issue needs more investigation (alongside the increased use of AHP in healthcare).

Related to the issue of trading risk, a concern that has been expressed in the literature concerning the SG technique is that, when asking individuals to trade, the attribute that is being traded becomes salient. That is, people give more importance to this attribute.[213] This has been shown within SG experiments in which individuals are first asked their probability indifference point between a gamble and certain outcome and are then presented with a pairwise comparison between these two scenarios (i.e. the certainty equivalent and the gamble to which individuals said they were indifferent). In such choice experiments the certainty equivalent is usually preferred to the gamble.[472,473] One possible explanation for this is that the attribute that is traded (risk in this example) is weighted more highly than in a situation where a choice is made.

Questions are therefore raised concerning whether individuals who are not willing to trade are protesting to the technique, are giving an exaggerated importance to the attribute being traded, or really do not value the health state being valued. This has not been addressed in the literature.

Perhaps the biggest challenge to the validity of the SG technique is the wealth of evidence showing that individuals consistently violate the axioms of EUT.[352] Brazier and colleagues[253] provide a review of this literature. This raises a question concerning the justification for the continued use of SG. The answer to this question may lie in the normative appeal of the technique. That is, violations of the axioms of EUT may be due to misunderstandings on the part of those making choices, and individuals may be willing to receive help from the theory to make optimal decisions. However, there have been increasing criticisms of EUT as a normative model of decision-making under uncertainty.[474–481] This is fuelled by the fact that, when asked to revise responses that violate the independence and reducibility axioms, many individuals refuse.[478] (Experimental evidence suggests that individuals will change preferences when alerted to intransitivities.[474]) Alternative theories of decision-making under uncertainty have thus been put forward. These argue that EUT fails not because individual behaviour is suboptimal but because the theory fails to take account of psychological factors of regret and disappointment.[476,477,479–481] Evidence has suggested that regret is an important element in individual valuation and decision-making in healthcare.[482,483] Questioning of the normative merit of EUT means questioning the utility estimates derived using the SG technique.

Time trade-off

As with the SG, the TTO method inherently provides the respondent with a constrained choice that elicits utility numbers that are held to represent strength of preference. The TTO technique was developed by Torrance and colleagues[247] specifically for use in healthcare research as a simple-to-administer alternative to the SG,[352] and a "less complicated, conceptually different although equally sound alternative to SG".[417] Like SG, the technique requires interview-based surveys so it will be relatively expensive. Whilst the technique has no well-defined theoretical basis, it does involve the concept of choice, and therefore may be argued to be rooted in welfare economics.

Brazier and colleagues[253] report good rates of acceptability.[363,462,463,484–489] For example, Detsky and colleagues[487] report a 97% completion rate, Johnson and colleagues[485] 100%, Dolan and colleagues[362] 95% and Glasziou and colleagues[488] 91%. Similar results were derived from the two additional studies identified in our review: Zug and colleagues[469] reported a completion rate of 94% and Handler and colleagues[490] 100%. However, as with the SG, it should be noted that such rates are for interviews so they would be expected to be high.

There is a limited amount of literature on the internal consistency of TTO responses. Brazier and colleagues[253] report Dolan and colleagues[362] as having a 92% consistency rate against 12 logically consistent comparisons. In addition, our review identified a study by Badia and colleagues[365] that concluded that the TTO technique produced the highest level of inconsistencies when compared to the VAS and a ranking method.

For the reproducibility criteria, test–retests are good for the TTO.[362,363,400,418,420,459,462,466,491,492] Brazier and colleagues[253] note that in comparing the test–retest measures for each of the health-status measures, the TTO probably presented the highest reliability.

In terms of convergent validity, Brazier and colleagues[253] highlighted reasonable correlation between TTO and SG valuations.[466–469,493] Brazier and colleagues[253] also cited Robinson and colleagues[494] where comparisons were made between the TTO and VAS. It was found that TTO valuations were a better reflection of health-state preferences than were VAS scores. Additional literature identified in our review by Zethraeus and colleagues[495] and Swan and

colleagues[496] both found that the TTO scores converged more with WTP than did rating scales.

One of the challenges to the validity of the TTO technique is in the techniques underlying assumption of constant proportionality. That is, the technique implicitly assumes that the amount of time an individual is willing to give up to be in a given health state is independent of the time horizon in which they will be in that health state. So, if an individual states that they would give up 2 years of a 10-year time horizon to be in a given better health state, then the assumption of proportionality implies that they would give up 4 years from a 20-year time horizon. However, it is possible that the value an individual gives to a certain health state is influenced by the amount of time they have to spend in that state.[49,462,497–500]

A further underlying assumption of the TTO technique is that individuals are willing to trade life expectancy. However, Brazier and colleagues[253] identified literature that indicated that individuals were unwilling to do this.[484,490,494,501] For example, Irvine[501] revealed that 47% of respondents did not accept a reduction in life span in order to obtain a shorter period in optimal health and Robinson and colleagues[494] reported that the TTO health states have a "threshold tolerability" whereby a certain period of time had to be exceeded before respondents would be willing to trade. As with the identification of non-trading in SG experiments, this may be a manifestation of the potential salience problem. As with SG, questions are therefore raised concerning whether individuals who are not willing to trade are protesting to the technique, are giving an exaggerated importance to the attribute being traded, or really do not value the health state being valued. As with SG, this has not been addressed in the TTO literature.

An additional study identified in our review found difficulties when applying the TTO technique to assess patient preferences for chemotherapy.[416] However, satisfactory results were obtained when the same analysis was performed with patients receiving hospital dialysis for end-stage renal failure. The authors note that "this is typical of TTO exercises as the approach is usually advocated for eliciting preference weightings for chronic health states … and while such patients may live with the cancer, or fear of its recurrence, for some considerable time, it is unlikely that any will be maintained on long term chemotherapy". This view is highlighted by Dolan and Gudex.[502] They state that the "TTO method is only feasible for valuing chronic health states that last for durations of five years or more".

Another concern for validity is found in Dolan and Gudex.[502] Here, evidence of framing effects were cited, arguing that values elicited using this technique may be a function of the questionnaire rather than the respondents' underlying preferences. Concern has also been expressed that the TTO presents individuals with an unrealistic choice, since the choice is framed in a certainty context.

Person trade-off[#]

Given the PTO's consideration of distributive issues, it is intuitively appealing.[251] However, it has been criticised for a lack of any theoretical basis.[49,250,503] Given that the choice question is phrased in a societal context, it cannot be seen to be rooted in economic welfare theory (a theory that has been linked to the TTO technique). However, in support of the technique, it has been argued to possess interval properties.[250]

Whilst very little empirical work has been carried out on the PTO technique, Brazier and colleagues[253] argue that the technique is promising, and future research should be carried out. The evidence presented by Brazier and colleagues indicates that if the PTO technique is to be successfully used within healthcare, it should be within an interview format.[248,253] For example, Nord and colleagues[504] report response rates of 28.2% and 27% for self-administered questionnaires and Nord[251] notes that PTO responses from 17 of the 53 respondents showed an unwillingness to take part in the PTO task. He also notes that the technique can be cognitively demanding, subject to framing effects, and inappropriate for use with self-administered questionnaires. Ubel and colleagues[364] also report difficulties using the PTO via a written survey, reporting high levels of inconsistency.

Brazier and colleagues[253] note that there is currently a lack of information on the reliability of the technique, and very limited information on validity. Patrick and colleagues[248] found

[#] Given that our review identified no additional studies, this section presents a summary of two reviews, that by Brazier and Deverill[49] and an update of this by Green.[250]

some evidence of convergent validity with respect to PTO and VAS, and Froberg and Kane[352] comment on a study that found the VAS and PTO to give similar results.

Willingness to pay

WTP has been widely used in healthcare to elicit public views.[256–258] The application of the WTP technique to healthcare has focused on using the PC and CE approaches.[259–263]

WTP has its theoretical basis in welfare economic theory.[505,506] Pauly[506] observed:

> "In welfare economic theory, there is only one accepted way to measure the benefits an individual gets from a program. Benefit is assessed as the willingness to pay for the programme when supplied with information as complete as it can be."

Given that benefits are measured in money, it is held to represent a cardinal measure.

The different approaches to WTP have revealed different levels of acceptability to respondents. The OE approach has prompted large numbers of item non-response and protest answers.[257,261,507–511] In comparison, the PC and CE approaches have demonstrated higher response rates,[261,367,506,512,513] lower item non-response[260,261,263,514–517] and less zero responses.[512] The bidding game has also given high response rates in an interview setting where it is best suited because of its complexity rather than a postal questionnaire.[353,518] Whilst postal questionnaires are considerably cheaper than interviews, they generate lower response rates. Flowers and colleagues[519] demonstrated the viability of automated computer-based questionnaires. There is some evidence of low response rates in older age groups.[470]

There is much evidence through direct questioning of respondents and interviewers that respondents have few problems in understanding the WTP questions.[505,507,519–523] On the other hand, there is also some evidence that respondents find the WTP questions difficult to answer,[355,524] and concern has been expressed that individuals will object to the technique on the basis of the technique being related to ability to pay. In the most comprehensive WTP project in healthcare, Donaldson's[354] six country European study found the WTP method to be "feasible", with only one country having more than 10% protest responses. However, Ryan[517] noted providers' opposition to the WTP technique.

One way of measuring the consistency of WTP is to ask respondents their preference for the commodities being valued and then compare such preferences with their stated WTP. A respondent that gives a smaller WTP value for their preferred option or greater WTP value for their less preferred option may be considered inconsistent. Studies using the WTP technique have shown inconsistencies between WTP values and stated preferences.[355,367–369,417,525] For example, Ryan and San Miguel,[369] in looking at women's preferences in the treatment of menorrhagia, found that whilst women stated that they preferred conservative treatment to hysterectomy, they were willing to pay more for the latter. In this study, cost-based responses were found to partly explain these inconsistencies. A number of other WTP studies have found evidence of cost-based responses. Schkade and Payne,[526] in using verbal protocol analysis to investigate how people respond to WTP questions, found that individuals justify their responses by referring to the cost of the commodity being valued. Donaldson and colleagues[355] found evidence of cost-based responses in a study looking at different methods of testing for cystic fibrosis. They argued that this problem may arise when OE questions are used to elicit WTP values. However, the study cited by Ryan and San Miguel found evidence of cost-based responses when using the PC.[369]

A number of studies have found evidence of good test–retest reliability,[354,419,519,527,528] although others have found the opposite.[417,508] Klose concluded that reproducibility remains an important area for future research.[257]

Whilst some studies have found strong indications of internal validity,[257,470,529–531] others have shown mixed results[263,355,368,465,507–510,518,532–535] or opposite to expected results.[367,511,536–540] Regarding convergent validity, Klose noted evidence of correlation with other measures of health benefits found in some but not all studies.[419,495,520] These include rating scales,[417,495,496,513,541,542] SG,[417,419,470,505] TTO,[417,495,496,505] and choice-based CA.[194,199]

Two studies were identified which had tested the criterion validity of the WTP technique. Granberg and colleagues[543] found that 55% of couples gave a WTP of more than £10,000 for *in vitro* fertilisation (IVF) treatment, which equates to the true cost of this non-NHS treatment (it should be noted that IVF is available on the NHS but access is restricted). Walraven,[544] in an empirical study in Tanzania, found a majority of patients actually paid more than they said they were willing to pay: 62% at one

hospital and 67% at another. At another hospital, which later introduced user charges, WTP predicted behaviour "reasonably well".

Concern has been expressed concerning the validity of the WTP technique for a number of reasons. One of the greatest challenges concerns the evidence of the insensitivity of WTP to the size of the healthcare intervention being valued.[262,354,356,509,530,533,545–551] However, a few studies challenge this.[552–554] Whilst no evidence has been found of question order effects,[354,553] there is evidence suggesting sensitivity to the payment vehicle,[548] as well as the technique used to estimate WTP.[354,555,556]

The WTP approach used may determine the extent and type of any biases found. Several studies detected the presence of 'yea-saying' with the CE approach¶,[263,356,555] some evidence of range bias with the PC[355,506,557] and starting point bias with the bidding game.[353,356,552,558] Further research has shown the results of the CE approach to be sensitive to methods of analysis, including treatment of 'don't know' responses, limits of integration and bid vector design.[559,560] Evidence has also been found of strategic behaviour.[532,544]

Measure of value

Given the obvious lack of use of this technique in healthcare, there is a limited amount of methodological work. The technique would obviously need to be carried out within an interview, so could be potentially costly. It may also be a lengthy and tiresome process if there are many alternative options to compare. Dickinson[265] considered the investment of 15 hours of management time involved as a "good buy". However, it has a number of attractions, not least the attempt to achieve the optimal bundle within the resources available. To this end, the technique incorporates the important notion of scarcity. The methodology of the technique would ensure internal consistency. There is no information currently available on reliability or validity, and research around these issues should be encouraged.

Allocation of points

Whilst little work has been carried out looking at the budget pie, the characteristics of the technique suggest that it could be potentially useful for setting priorities. Given its obvious ease of use,[561] it is relatively inexpensive to carry out (because mailed questionnaires can be used).[282] Whilst a theoretical basis for the technique has not been developed, individuals are explicitly forced to think about trade-offs and strength of preference. Indeed, Clark[277] claims that given that the technique uses money (or points) to generate trade-offs, the resulting measure is cardinal. The case for this is strengthened by the fact that respondents are told to think about their strength of preference when allocating money/points. Clark[277] reports unpublished work by Ostrom which found high response rates across educational backgrounds, whilst Strauss and Hughes[282] reported a response rate of 28% (1001/3517) from postal questionnaires. Test–retest reliability was declared as "modest" by Srivastava and colleagues[562] (their questionnaire included Likert and ranking questions). Clark[277] states that although the budget pie method is tentatively appealing, results depend on conversion of money to utility, a direct relationship between money and utility, and honest revealing of preferences. Convergent validity between the budget pie and Likert and ranking questions have been shown to be poor.[562] Mullen and Spurgeon[563] in their review of the budget pie technique and its use in the healthcare noted that scaling problems could be encountered when using the budget pie literally, i.e. dividing or allocating monetary budgets. If the budget is going to be realistic, the ability of respondents to cope with sums that are perhaps beyond their normal experience is compromised. By the same token, using unrealistic sums may result in unrealistic responses. This has a bearing on whether realistic or 'honest' preferences can be elicited, which in turn may mean that cardinality is not maintained and intensity of preferences is not accurately reflected in individuals' responses. It has therefore been more common in recent studies[564] to define a 'budget' in terms of a set of tokens or points that must be divided or allocated across alternatives.

Instruments for eliciting preferences

Allocation of points

The PGI has been assessed against the short-form 36 (SF-36) for convergent validity[269–271] and been shown to perform reasonably well. It is reproducible for group analyses but not for comparison of individuals as assessed by test–retest.[269–271] It elucidates relevant items on a much wider range than in pre-set techniques.[272,273] The concern with the PGI is its acceptability. Correctly completed response rates are low: of a response

¶ Yea-saying refers to the tendency to say yes to whatever you are asked.

of 75.4% in the original study of back pain, only half completed it correctly,[269] and Macduff and Russell[270] reported similar findings. Thus, while it has the advantage of offering individual choice, it is not yet usable in practice, especially as a postal tool for prioritisation.[270]

The SEIQoL–DW is reported to be more comprehensible and acceptable than the original SEIQoL or PGI, but it is interviewer-delivered by clinicians or researchers (both of whom like it[274,565]), who may therefore be able to ensure understanding. The two methods are not interchangeable because they are conceptually different.[565] As noted above, SEIQoL elicits implicit judgements whereas, by definition, DW is explicit. However, the authors use the former when there is no time constraint and the participants are not cognitively impaired.

Discussion and conclusion

The systematic review identified a number of quantitative techniques that have been used for eliciting public preferences. Techniques identified were classified as either ranking exercises, rating exercises or choice-based exercises. An important distinction was made between instruments and approaches for eliciting preferences. Conclusions regarding the methodological status of techniques are stronger for the former.

Ranking exercises included simple ranking questions, QDP and CA. The popularity of simple ranking exercises probably reflects the ease of both devising a ranking exercise and analysing the resulting data. Whilst it is recognised that ranking exercises may provide policy makers with some useful information, their failure to incorporate any concept of opportunity cost, lack of any well-developed theoretical basis and lack of any measure of strength of preference, renders them of limited practical use when setting priorities.

The QDP has not been used to date in healthcare, and therefore there is very little empirical work assessing the methodological status of the technique. One of the potential strengths of this technique is its ability to deal with the vagueness that exists in human decision-making. This is potentially relevant in healthcare, where preferences may be both vague and difficult to articulate. Given its well-developed theoretical basis, and its strength of preference properties, future work eliciting public preferences should consider this technique.

CA (ranking) exercises are potentially very useful at the policy level. From such output it is possible to estimate the relative weights of different attributes/criteria in the overall provision of a good or service, the trade-offs individuals make between these attributes, and an overall benefit or utility score for different ways of providing a good or service. The ranking approach is attractive in that it has a well-developed theoretical basis, and cardinal data can be obtained from ordinal data. Whilst the ranking approach has received little attention in healthcare (relative to the choice-based approach), future work should explore the use of this technique.

A number of different rating scales, and applications, were identified. The VAS has proved popular as a method for estimating quality weights within the QALY paradigm. Whilst the technique is attractive in terms of the relative ease with which quality weights can be estimated, its main limitations are the lack of any constrained choice and the doubts expressed over whether the technique measures strength of preference. Given the availability of alternative techniques, the suitability of this technique for estimating quality weights must be questioned.

CA rating scales have been applied in healthcare, and the limited methodological work available indicates that the technique did well against the predefined criteria. As with ranking exercises, one of the attractions of this approach is the elicitation of cardinal measures of benefit. Future work should explore the possible uses of this technique. A variation of CA rating scales is the SEIQoL, a technique used in the health outcome literature to estimate quality of life scores for patients. There is currently a lack of empirical work on the methodological status of this instrument. However, given it is essentially the same as CA rating exercises, future work should be encouraged.

A number of techniques have been applied by social and behavioural scientists to elicit attitudes. Likert scales are most commonly used, but studies were also found which had employed Guttman scales and the SDT. Whilst such scales provide useful information, it must be recognised by users that there is no strength of preference measure, nor any consideration to the importance of the different components that made up the score.

Satisfaction surveys have been frequently used to elicit public opinion. Their popularity probably reflects the relative ease of carrying out such

surveys. Researchers should ensure that they construct sensitive techniques, or else use generic instruments where validity has already been established. Even when well-designed, sensitive techniques are used, users should be aware of the limited use of such techniques at the policy level. SERVQUAL appears to be a potentially useful instrument and future research should consider its application in healthcare.

A number of choice-based techniques were identified. Three such techniques – MoV, AHP and allocation of points technique – have had limited application in healthcare, resulting in a small literature on their methodological status. However, they all appear potentially useful. The MoV technique, although requiring an interview setting, considers the optimal allocation of resources within a given budget. AHP has a sound theoretical base and incorporates intensity of preference. The apparent simplicity of the allocation of points technique, together with its explicit attempts to consider trade-offs and strength of preference, suggest that it should form the focus of further research. Whilst the SG technique has been argued to be the gold standard as a method for eliciting utility weights within the QALY paradigm, concern is expressed over the evidence suggesting that individuals consistently violate the axioms on which it is based. TTO, discrete choice CA and WTP all did well against the predefined criteria and should continue to be researched. More methodological research is required on the PTO.

A number of techniques have been identified above for both applications in healthcare and future research. A priority area of research is to address the cognitive strategies and decision-making heuristics respondents adopt when completing quantitative surveys. This should involve extensive qualitative work to inform the design and interpretation of quantitative studies.

Chapter 6

A review of qualitative techniques for eliciting public views

This review concentrates on qualitative methods that have been, or could be, used to elicit public opinion regarding priority setting in healthcare. The qualitative techniques identified in chapter 3 are reviewed according to the criteria in chapter 4. Techniques were classified as either individual-based or group-based. It was recognised that the terminology of the chosen criteria would not have been used in most existing reviews of qualitative methods, and that therefore there would be an element of interpretation or translation in our review*. Given the nature of qualitative research, all methods identified may be seen as approaches to eliciting preferences (as opposed to instruments).

This chapter sets out techniques in sequence. For each technique general issues regarding the methodological status of the technique are first discussed. Secondly, where identified, examples are given of how the technique has been used in primary research, and methodological issues are highlighted in respect of the chosen criteria. In this second part, papers have been included in the review only if they adhere to the inclusion criteria set out in the methods section (chapter 2); that is, if the author/authors conducted a primary piece of research and made some attempt, either explicitly or implicitly, to evaluate it. Where a comment about a criterion has been made explicitly by an author/authors it is included here only if it is consistent with our terminology. Where a comment is made implicitly, that is if it fits in with our criteria but the exact word is not used, it has also been mentioned, but is referred to using our definition of the term.

Individual approaches

One-to-one interviews

General methodological issues

The methodological literature reveals that there are several strengths to the interview method. The interviewee has the opportunity to control the direction in which the interview is heading so that true feelings can be determined, thus increasing the potential validity of the technique.[381] Because the interviewer is present ambiguities can be clarified and misconceptions checked at the time of the interview.[9,381] Interviewees are offered sufficient time to express their ideas and have the opportunity to ask questions and have these fully answered.[307] They provide those people who are not particularly vocal in public settings, in a public meeting for instance, with an opportunity to offer their opinions about a certain topic. "These may offer a way of enabling the silent voices to be heard."[1] Ultimately, "the only valid way of hearing the voice of the public may be by one to one interviews".[566]

However, weaknesses have also been noted. Respondents may give socially desirable responses, wishing to please the interviewer and expressing ideas that they think the interviewer is seeking.[567] There is a general problem with interviews in that they are not an anonymous research method[351] and therefore some people may withhold information, particularly if the topic is very personal or embarrassing. Respondents might be looking for confirmation from the interviewer in response to their answers. There is, of course, also risk of interviewer bias.[371] Objectivity is another problem, given that the researcher can never be truly objective; this should be recognised and steps should be taken to minimise these factors.[9]

The method of analysis is important for reliability, validity and objectivity. Interviews may be recorded and transcribed and recurrent themes identified,[384] or the interviewer can simply make journalist-type shorthand notes during the interview. The decision regarding recording interviews should take the concerns of the interviewee rather than those of the interviewer into consideration, but, if the two differ, neither is regarded as more valid, reliable or objective than the other.[384] It is widely believed that displaying statements verbatim helps to give the reader a feel for the kinds of comments made in

* Given this, the evaluation of qualitative instruments may be more open to problems of inter-observer reliability.

the respondent's own words. There are, however, a number of problems associated with this format. There is a danger that quotes can be taken out of context or become distorted.[9] Cost is another consideration regarding conducting and transcribing interviews. It is suggested, for example, that 4 hours or more be allowed for the transcription of every 1 hour of recorded material[568] and the reporting of the findings can take 4–6 weeks.[379]

Examples of the application of the technique

Interviews have been very widely used in health-care research. Here several examples and the specific methodological problems encountered are presented as illustrations. Additional examples are highlighted that look at using the telephone, email and the dyadic format.

In the study by Ayanian and colleagues,[288] patients were selected from a stratified random sample of patients undergoing dialysis. They were interviewed by telephone about issues surrounding their treatment and attitudes towards it. One major drawback of interviewing in this study raised by the authors was that the sample obtained may not have been sufficiently representative of racial issues precisely because a portion of the intended sample was debarred from taking part because they could not speak English. By the criteria of this review, this concerns the generalisability of the results. Additionally, results achieved might have been due to the doctor who was advising the patients; that is, differences could be attributable to the doctors being of a different ethnicity to that of the patient. This would affect the validity of the study.

The study by Dicker and Armstrong[289] involved semi-structured interviews with 16 patients from an inner-city general practice. Steps were taken to compensate for the small number of participants by ensuring that a genuine cross-section of the community was recruited. Thus, representativeness was sought by deliberately selecting a range of different people using the service – a sampling technique known as purpose or **purposive** sampling.[8] This study found that respondents answered in relation to the context of the choices and considered a more general approach detached from their own motives. Here the criterion of acceptability was addressed: the authors reported that participants found it unacceptable to consider the choices from a personal perspective. This was perceived by the authors as a way for participants to appear unselfish, or to show a genuine concern for their friends or relatives.

In the study by Williams and colleagues,[290] 15 people referred to a community mental health team were interviewed both before and after their consultation to ascertain if they were satisfied with the treatment they had been given. It is not fully detailed in the text how this sample was selected and therefore the possibility of a biased sample cannot be discounted. Careful consideration was, however, given to the nature and setting of the interviews. It was felt that the people would feel more comfortable being interviewed in their own homes. The authors stated that interviews conducted outside of the medical setting would help to ensure that they would not take the form of a medical consultation. Participants were encouraged to talk about their experiences in lay terms; in the terminology of our review, this may also have helped to ensure that the questioning was acceptable to the participants which, in turn, is likely to increase reliability and validity. The analysis stage is detailed in the text. A grounded theory approach was adopted, and new patients recruited in order to provide further explanations for information deemed to be relevant to the study's aims. The second round of interviews was presented as a way of checking the validity of the study because participants were given the opportunity to provide feedback concerning the interviewer's perception of what they had stated in the first round. The study used a small sample of patients, which ensured that each case could be examined in detail, but has implications for the generalisability of the study.

Crabtree and Miller[291] found that long interviews can take between 1 and 6 hours to complete. Between three and six OE questions are asked and are designed to draw out long, detailed answers. The development of the questionnaire was carefully conducted in this study. Issues of validity and reliability were considered at all stages; for example, all of the members of the research team listed their beliefs, derived from personal experience, about pain. This was done in order to understand biases that might be introduced. Interestingly, at this stage the authors mention that biases cannot be eliminated but can be recognised and used effectively in the research project. The authors recognised that it was impossible for the researchers to be entirely objective, but took steps to comprehend and utilise the individual ideas brought in by the researchers involved. During the interviews a number of 'floating' and 'planned' prompts were used to explore further the responses of the interviewees. The authors stated that "if the

long interview is successful, the outcome will be an interpretation that simultaneously reflects the participant's reality and has generalisability to theory". Pilot interviews were conducted first to modify and improve the questionnaire, and secondly to improve interviewer performance in terms of reducing intrusiveness and distortion. A purposive sample was obtained to ensure that interviewers and interviewees were strangers, that respondents did not have a specialised knowledge of the topic under discussion and that there was a spread of age, gender and status. Four physicians and four patients with recent experiences of pain were selected. Three researchers read the transcripts and identified utterances, thus enhancing validity and reliability. These were then systematically compared and summarised into a quantitative style code-book. Each researcher identified emerging themes. The process of comparing the physicians and patients with one another was then carried out, although it was not completed at the time of publication of this review. It is stated that the conclusions drawn here were generalisable and transferable to other settings. It was concluded that this is a lengthy process; for example, 256 hours were required to analyse eight interviews. It is stated, however, that in total the project should be able to be completed within a year, which is comparable to "many primary care epidemiological studies" and therefore an acceptable time scale. Additional costs included interviewer time, transcription time and use of a computer.

One-to-one interviews have been compared with other forms of interviewing. Wilson and colleagues[292] discussed the merits and pitfalls of using face-to-face and telephone interviews. The paper highlights many of the issues that need consideration when embarking on a one-to-one interview. A study looking at continence care was undertaken. A pilot study used along with the five considerations presented by De Vaus (cited in Wilson and colleagues[292]; see appendix 3) revealed that telephone interviews would be the most appropriate technique for use in the main study. Considerations included response rates, representativeness of the sample, design of the interview schedule, anonymity and cost. From a practical point of view telephone interviews were perceived to have several advantages, including providing greater interviewer safety. Indeed, it was expressed in the text that on some occasions the interviewee would become 'over familiar' with the interviewers who were relieved to be conducting the interview over the telephone. Also, it was found that four to five, and sometimes six, interviews could be conducted per day, compared with two or three face-to-face interviews. This can help to reduce interviewer travel and interview time costs. Groves and Kahn (in Baker[8]) found that each interview worked out to be less than half the cost if it was conducted over the telephone. The reduction in time was seen as being advantageous because fatigue may be an issue if the interview is very lengthy, which may affect the quality of answers, which in turn may affect reliability and validity. Such interviews are more impersonal than face-to-face interviews and therefore the respondent may be more willing to talk about personal or embarrassing issues. It was perceived by Groves and Kahn, however, that respondents may feel slightly rushed and less relaxed than if the interview was face-to-face.[8]

Telephone interviewing might use slightly different language and more clarification or explanation may be required because there is no opportunity to present visual cues.[8,9] In addition to this, responses may be shorter and less information divulged over the telephone because probing is easier in the face-to-face situation.[8] This may challenge the overall reliability and validity of the technique.

It has been proposed by Foster[287] that using the Internet and email services may be a practical alternative to face-to-face interviews. This technique was used to look into how teachers and lecturers plan and design their courses. Foster stressed that this is a very new technique and one that needs much consideration and careful thought before employing it as a research method. The main advantage of such a medium is that time, and therefore money, can be saved owing to the elimination of transcribing interview tapes. The information arrives in a form that can already be recognised and analysed by the computer, meaning that transcription is not necessary. In addition, this method can be logistically advantageous. There is no need for time to be spent travelling to participants' homes or offices, participants can reply at their own convenience and can decide whether to take part or not without feeling pressured in any way. However, there are a number of disadvantages to the method. There is a risk of it being viewed as 'junk' or 'nuisance' mail, and only those with a specific interest in the subject matter may reply; this has implications for acceptability, validity, reliability and generalisability. In addition, there is no opportunity for the interviewer to probe new issues coming up, or to clarify points misunderstood by the respondent. Identifying

suitable audiences is also a problem, and using this method alone could miss a large portion of important opinion; therefore it is concluded that this method is unproven.

The dyadic interview

General methodological issues

Sohier[293] argued that this type of interview maintains objectivity, enhances the credibility of the data, and therefore the reliability, and is also ethically sound. He argued that if two closely related people are interviewed together then more detailed and contrasted information can be liberated. Information can be clarified or contradicted, thereby providing a more reliable and valid set of results. He noted that if the interviewee becomes overwhelmed with emotion because of the difficult topics under discussion, the second participant can then take on the role of comforter and can take over the dialogue for a period of time. This aids the interviewer as it enables him or her to "maintain the research posture" and can enhance the quality of the data. The situation may occur where both parties become very upset and emotional. In this instance the interviewer must decide if it is appropriate and ethical to continue, or to close the interview. This highlights again that the interviewer must be highly skilled. Another disadvantage suggested by Sohier is that sometimes people may not want to reveal too much information in front of those very close to them. He stressed that confidentiality must be assured and that the interviewer must remain compassionate and non-judgmental throughout the process.

Case study analysis

General methodological issues

Given that case study analysis represents a new method of analysing qualitative data, there is very little methodological work considering this technique. However, it has been noted that whilst case studies have the advantage of providing a set of extremely rich data, this is at the expense of generalisability.[569]

Delphi technique

General methodological issues

The Delphi technique was reviewed in a recent HTA report by Murphy and colleagues.[570] An advantage of the Delphi technique is that it is often difficult to organise meetings with professionals or 'informed individuals'.[570] There is no need for participants to meet up in one location and therefore a wider scope of opinion can be gained because the logistical problems of organising meetings are eliminated. Also, as the participants are consulted on a number of occasions there is the opportunity for statements and suggestions to be changed or withdrawn as a period of 'considered thought' is allowed. The anonymity of the process could be extremely advantageous. It is possible that more controversial issues could be raised if individual identity is protected. The criterion of objectivity is rarely mentioned in the literature. However, Williams and Webb[571] stated that researcher bias can occur when analysing the findings. Murphy and colleagues highlighted that there has been little research in assessing the quality of the Delphi method.[570] However, in several of the studies mentioned below an attempt has been made to evaluate the technique and, in some cases, the study itself.

The technique also has a number of disadvantages. Response rate can be a problem and often decreases as the rounds progress,[571] leaving the method open to response bias. The Delphi technique is a consensus method and this may mean that some important ideas are eliminated.[303] These issues affect both validity and reliability.

Examples of the application of the technique

In the study by Guest and colleagues,[294] they cited Le Pen and colleagues as stating that this is frequently used in healthcare assessment and is "a useful tool in situations where data are not available, but resource allocation decisions need to be made".

Endacott and colleagues,[295] in their study on critically ill children, carefully evaluated and reflected upon the limitations of both the technique and the application of it. It was found that, although attempts were made to minimise the workload of paediatric intensive care sisters, it was still a considerable amount, thereby affecting the acceptability to respondents. This may also have affected response rates and therefore gained only the opinions of a limited number of professionals. In 1994, it was said that there is no evidence of reliability for this technique.[571] In the application presented here, however, extensive efforts were made to maximise validity and reliability by piloting each round on nurses who worked with critically ill children.

Harrington[296] found in his study on senior practitioners on research priorities in occupational health that, although a degree of consensus was reached, it was remarked that perhaps asking such a specific group to comment could limit the scope of opinion; thus, the representativeness of the panel was questioned. It is also stated that the

types of professionals included here may be concerned with strategic as opposed to practical issues, limiting the study in terms of reliability and validity. Charlton and colleagues,[297] in their study on spending priorities of different health professionals, found evidence to suggest that members of different professional groups do not change their opinions after obtaining feedback relating to other groups, therefore strengthening the argument for having a multidisciplinary team involved. A flaw in the validity of the method can be seen when one of those involved was asked to comment on the results: she felt that her priorities stated in the questionnaire were not necessarily her choices for priorities in research, which suggested that the technique required modification.[296]

One study carried out by Roberts and colleagues[298] used a combination of methods and compared the responses of consultant geriatricians and patients. Using the Delphi method, 89 geriatricians were asked to formulate a list of performance measures. Also, 44 day-hospital patients were interviewed and asked to state those factors that they considered to be important to them. It concluded that, as differences were found between the priorities of the two groups, it was advisable to involve different interested parties when making priority-setting decisions. The information was gathered from the patients by interview; while this could have affected the sample size and therefore the generalisability of the study, it was seen as advantageous but also influential to have a researcher present for the patients should they have any queries.

When considering the criterion of cost it is said to be a relatively inexpensive method of gaining a large number of responses.[9,571] This is especially true because questions are distributed by post, which additionally ensures anonymity from other participants in the study. This is stressed as being an advantage of the Delphi method as the issue of a few dominant individuals taking over discussions will be avoided.[297] However, Charlton and colleagues[297] experienced non-response as a result of minimising costs. It was decided that in order to save money that no initial meeting be held explaining the study to those targeted; this meant that there was a lack of knowledge about the objectives of the study and in turn led to a large number of invitees not participating.

Hadorn and Holmes[299] used Delphi to ask surgeons, clinicians and "relevant specialists" their opinions on elective surgical procedures. However, little information was provided in the article on the inclusion criteria and numbers involved. Also, to include social factors in the criteria two public hearings were held. Members of the public were selected randomly but it is not made explicit how this was done, how many people were involved and how these factors would affect the suggestions made and points raised at these meetings. These factors combined make it very difficult to evaluate the use of Delphi in this application as far as reliability, validity and generalisability are concerned.

Gabbay and Francis also conducted a study using more than one group.[300] The Delphi technique was applied to ask general surgeons and anaesthetists their views on the maximum potential for day surgery. There were problems with generalisability because the surgeons involved were reluctant to comment on subspecialist areas, questioning the reliability of these parts of the study. Cost is not mentioned but the method appeared to be acceptable to the respondents as a large proportion participated in all stages.

Wilson and Kerr[301] asked 353 bioethics society members and their designates about important social values related to healthcare. However, generalisability is questionable because the sample obtained was not representative as it consisted of well-educated, affluent members of society only. Missing out those in lower socio-economic groups and those with less knowledge of healthcare seriously limits the study. Additionally, the technique involved four stages. The fourth stage was sent to non-responders to the initial phases and was used to establish reliability and generalisability to the results obtained earlier. This forth stage yielded a poor response rate and therefore detailed comparisons were not made. Thomson and Ponder[302] used the Delphi technique to develop a survey technique in order to identify priorities for the Texas Society of Allied Health Professionals, along with other state affiliates. A panel of five people chosen by the researchers nominated 21 health professionals to take part in the development of the survey technique. This non-random selection of participants could have resulted in recruiting a biased sample. Time-consuming efforts were made to trace those who did not respond and the method was piloted to establish a Delphi question for use in the main study that was neither too vague, nor too narrow. The statements made that were subsequently eliminated were reported and reasons for their exclusion provided. Based on the researchers' experience of applying Delphi, a list of seven suggestions was provided to guide the future development of a survey technique.

One study that involved patients in the Delphi process was conducted by Gallagher and colleagues[303] who looked at policy priorities for improving care for diabetic patients. In all, 28 'experts' were questioned, including patients who were already receiving treatment for diabetes. A high degree of consensus was found. Reliability was addressed by using clear selection criteria and attempting to contact non-responders. Researcher bias was minimised in the collation stage by all three authors categorising responses independently. The researchers were aware of the potential loss of issues through centralising opinions.

Another study that attempted to incorporate the views of the public was carried out by Burns and colleagues.[304] Representatives of relevant organisations were invited to contribute, although it is unclear how many participated and in what capacity. The study concerned building consensus on the care of those with physical disabilities. Reliability checks were made for correlation of results for rounds two and three by comparing issue-by-issue point scores in these two rounds. It was stated that these showed a high level of correlation and were therefore very reliable. A second reliability check was undertaken by looking at the responses of people who joined the study at the beginning (stage 1) and those who joined later on (stage 2). Again, no significant differences were identified between these groups and the authors stated that the "addition of participants did not significantly alter the study".

Kastein and colleagues[305] looked specifically at the issue of reliability of Delphi in a primary healthcare setting in The Netherlands. They conducted a study to develop evaluation criteria for the performance of GPs when they are consulted about abdominal pain. A secondary objective was to look at the reliability of their application of the technique. They deal with "minimising situation specific biases" (this terminology refers to what we have indicated as quixotic reliability) by standardising recruitment procedure. They also sought to minimise "person specific biases" by carefully selecting their sample. Both family physicians and specialists were invited to participate. However, as previously discussed, merely obtaining a reliable sample does not ensure the reliability of the technique as a whole. Diachronic and synchronic reliability were not dealt with, as well as other issues concerning validity, objectivity of researchers, generalisability and cost. This study is, however, one of the few that attempts explicit evaluation.

Complaints procedures

General methodological issues

This type of exercise, although useful in identifying complaints and comments, will only incorporate the views of those with particularly strong opinions and those with especially negative experiences, so that generalising the results of this type of study is extremely difficult, if not impossible.[572]

A number of additional studies looked at using this method for obtaining public preference. Hemenway and Killen[573] suggested that using the complaints of dissatisfied patients can be used as an inexpensive and effective way of discovering information that can be used by policy-makers, but warned that this information should be used wisely because all complaints are not necessarily valid.

Reid and colleagues[335] discussed the role of the complaints procedure within the NHS with a view to quality assurance. The 'Oregon experiment'[574] looks at involving the public in healthcare rationing. Using complaints can indicate the strength of feeling concerning a particular treatment or intervention, but again, these complaints may be invalid.

Dean and colleagues[575] highlighted that there are two different types of complaint. The first concerns dissatisfaction with a service that has been provided and the second is concerned with refusal of treatment or intervention. Using complaints as a rationing technique was described as being dangerous because of the possibility that "services will go disproportionately to those who are most determined, articulate or litigious" (p. 343), which raises serious questions of equity in healthcare.

Group approaches

Focus groups

General methodological issues

Focus groups are sometimes perceived to be an easy method.[576] It is demonstrated here, however, that it can be a very challenging method to use and needs careful consideration and planning. A number of general principles must be considered when undertaking this method of research. Having a trained moderator/facilitator is vital for good focus group research.[307] The quality of information gathered can be determined by the expertise of the facilitator/moderator;[576] for example, expertise in skilled questioning and careful

probing. The facilitator is given the choice of how much control to have, but a certain amount of free-flow is needed and therefore careful attention needs to be paid to the selection of the moderator. It is not always the moderator's role to find common ground but to identify different and perhaps conflicting views from group members.[577] If the facilitator is not suitable, the results of the group may be misinterpreted and inappropriate topics discussed[331] and bias may be introduced; Festervand highlighted that this can affect reliability.[306] Participants are sometimes asked to take part in games: sorting vignette, "specially designed board games, 'cake cutting' or coloured disc games, for example, in groups about health priorities and allocation of resources".[9] This indicates the use of quantitative techniques, as well as qualitative techniques, within focus groups. Another important consideration is the setting. It is vital that participants feel relaxed, and that the room is comfortable. Additionally, it must be carefully thought out as to whether it is advantageous for the group to be friends or strangers because this can have implications on the depth and kind of information disclosed. These issues will be discussed using a number of examples below.

A major advantage is that group interaction can be observed.[307] The interactions differ by whether or not participants know each other; for example, people who work together or see one another on a daily basis, may feel inhibited when discussing personal or work-related issues. Group size can have an effect: Tang and Davis[578] listed the following critical factors in determining optimal focus group size: aims of the study; the number of question asked; the allotted time for each question; the format of the focus group session and the duration of the session.[578] Combined group effort produces a wider range of information and ideas to be expressed compared with one-to-one interviews.[306] Group discussions can be stimulating and members can find security in numbers and will be more willing to divulge information. Also, as participants are not required to answer each question, the responses are likely to be more meaningful. In addition to this, one comment can stimulate others into thinking of related ideas producing a snowballing effect and therefore perspectives that may otherwise be missed can be explored.[579] This technique gives people the opportunity to focus upon common rather than individual interests.[311] The discussion process in itself enables participants to form opinions, armed with fresh knowledge. People often do not know what they think about certain issues, especially those in which they have little prior knowledge.[311]

Additionally, they are perceived as being easy to conduct[309] and yield quick results.[310] In reality this is not the case, conducting a focus group takes a lot of planning and organisation, and requires skilled professionals to conduct them. In comparison with individual interviews, larger sample sizes can be used and it is possible to explore issues in depth and pursue unanticipated areas.[310]

There are a number of disadvantages associated with focus groups. Although attitudes can be expressed, the extent of these attitudes is difficult to gauge.[576] The logistics of arranging and carrying out a focus group can be tricky: getting the participants together in one place at one time can present problems.[307] Focus groups can be conducted quickly and more cost-effectively than other methods such as surveys, which means that inappropriate topics may be explored.[306] However, relative to a questionnaire survey it is still a costly exercise because transcription and reporting are time-consuming and thus labour intensive.[379,568] Individual responses and opinions can be influenced by the rest of the group.[309] Some people may respond in a way deemed to be socially acceptable, therefore swaying the results and affecting validity. This can be especially true if investigating a sensitive or controversial topic.[379] People may find it unacceptable to discuss these types of issues in front of or within a group of people, and therefore not reveal the private issues that they would on a one-to-one basis.[307] Compared with one-to-one interviews, there is less time for each person to talk and thereby to reveal their preferences.[307,576] In addition to this, some ideas may not be expressed because of concern that others will perceive them as socially unacceptable, racist or sexist. Participants may also experience pressure towards consensus and unanimity.[580] It is recommended that, to ensure reliability, two researchers independently code and analyse the transcripts, compare them, and agree upon important themes. This can also help to ensure that objectivity of the researcher is achieved.[9]

Because of the intensive nature of the method, sample sizes are small and therefore findings may not be generalisable. An aid to overcoming this is the careful selection of members that can represent the study population as closely as possible. All too often, however, the sample group proves not to be representative.[306]

Additionally, issues of self-selection or self-exclusion need to be addressed. It is likely that those who are particularly interested in the topic being studied have particularly strong opinions, or those who are extrovert are more likely to volunteer than those who are shy or withdrawn.

Methodological issues involved in using focus groups in studies in healthcare suggest that in some instances it is appropriate to 'break the rules' of conducting focus groups, emphasising that each topic needs to be individually considered and that it is difficult to adhere to a predefined formula.[313]

Examples of the application of the technique

Somerset Health Authority used focus groups to involve the public in priority setting.[311] Eight focus groups of 12 people each met three times a year. Professional recruiters were employed to gain a representative sample. The CHC members were used to test the effectiveness of focus group discussions. Groups discussed real issues rather than hypothetical ones; for example, the question of allowing and funding treatments outside the county was addressed. Background information was supplied and options of a figure to be agreed upon suggested. A briefing paper was also sent to participants in advance of each meeting and discussion was encouraged with friends and family members – something that was undertaken by some group members. The sample population was demographically representative of the community but with a slight bias towards those with a greater experience of using the health service. The authors argued that using professional recruiters made the focus group representative and that trained and experienced group facilitators led to the creation of a "richer bank of information than could be obtained through a more conventional structured survey".[311] The analysis was not made explicit about, for example, who was involved and what methods were used. Had this been done the validity and reliability of the technique could have been further established.

Kuder and Roeder[312] investigated public attitudes towards age-based priority setting in America. Five groups were held and separated into homogenous groups based upon age and socio-economic status. Gender and race were not considered in the selection process, which could therefore have influenced the results; for example reference to race was avoided in the groups. Two researchers who identified and agreed upon general themes carried out the analysis of the transcripts. Additionally, the coding process was tested for reliability by asking another member of staff to code one transcript. It was found that participants of all ages were reluctant to endorse a policy that limits care to the elderly, and were unable to decide the extent to which the government should be responsible for subsidising treatment costs.

In contradiction to what Morgan[309] and Krueger[310] said, there are situations where it is appropriate to use focus groups to explore sensitive or embarrassing issues. This is demonstrated in the Cohen and Garrett[313] study where client/worker relationships in residential mental health settings were looked at. During the focus groups the facilitator found that issues such as physical abuse and suicide attempts could be discussed amongst group members. They would support one another and offer information on similar experiences. It was important for the facilitator to gain trust before proceeding, and it was reported that this trusting relationship enabled in-depth data to be gathered. This again highlights the necessity for the moderator to be an experienced group interviewer.

Kitzinger[314] also investigated a sensitive topic: 52 groups were conducted, comprising 351 people, to investigate attitudes towards and knowledge of AIDS. The groups were selected from pre-existing social or work-based groups. These 'natural clusters' were chosen for a number of reasons. First, they were more likely to challenge one another because they had knowledge of each other's circumstances and, secondly, discussions could take place in social contexts that could aid in increasing the reliability of results. The facilitator engaged the participants in exercises such as card games to encourage everyone to talk, which helped the shyer group members to contribute at the start of the session, thereby boosting confidence for the rest of the discussion. Kitzinger was particularly concerned with maximising group interaction stating that this enabled unexpected issues to be covered through sharing common experiences and arguing over issues, so that both similarities and differences between members could be examined. The purpose of the study was to find diversity rather than to establish representativeness. Results, therefore, may not be generalisable, but nevertheless the study provides the reader with an in-depth insight into the views of different groups. Because of the size of the population this study must have been extremely costly, which is one of the main drawbacks of this type of data collection.

These two studies can be used to highlight that in some situations discussing sensitive or

embarrassing topics is acceptable to participants and that, again contrary to Morgan[309] and Krueger,[310] rich, in-depth data can be extracted. This may be because group members feel that their participation is legitimised and valued. It is possible that people can feel intimidated if asked to talk in front of others who are perceived to be on a higher social or professional level.

Ward and colleagues[315] found that certain topics are not suitable for discussion in focus groups. This was found to be the case when male attitudes towards vasectomy in Latin America were sought. It was unacceptable for men to admit that they had had a vasectomy as it could be associated with being "less of a man" and therefore inappropriate for public discussion. This can be used to illustrate that the way in which the question is asked can be very important to the reliability and validity of results. It was found that responses were provided to questions about whether having a vasectomy would affect their status in the community, rather than being asked directly about personal experiences.

In the study by Carey and Smith,[316] it was found here that using focus groups was unacceptable. This was because topics such as sexual orientation and drug and alcohol use needed to be addressed. Discussing these topics with colleagues was not possible for the members because disciplinary action might have been taken.

Another example where the 'rules' of conducting focus groups were broken is in a study carried out by Powell and colleagues.[317] Both users and providers of mental health services were asked to discuss provision of and access to these services. The groups were pre-existing, and this was seen as advantageous as the familiarity of members would provide a supportive atmosphere. Staff from different professional backgrounds were chosen in order to "maximise the discussion of different perspectives" and were of an equivalent level so as not to inhibit contributions by staff who may feel intimidated by the presence of higher level members. The groups were self-selecting, and no attempt was made to stratify for age, sex, race, occupation or length of contact with the service. The facilitator asked broad, OE questions and attempted not to lead the discussion. Reliability was enhanced at the analysis stage by using two facilitators and comparing their analyses. The authors stated that the focus group method was not used to its full potential; it was used here to provide validity for questions already being asked of users and providers in questionnaire format.

It was recognised by the authors that, aside from the limited generalisability of focus groups in general, this study has a circumscribed generalisability because the data collected only relates to the personal experiences of those in attendance, and that both the user and provider groups were not representative. Also, because the groups were pre-existing it is likely that they were composed of the more confident, vocal people, and would therefore miss the important inclusion of the quieter and perhaps more frequent users of the service. These limitations are recognised and acknowledged by the authors who state that the validity is not affected as they were merely attempting to validate pre-existing questionnaires and to generate new ideas for questionnaires.

It has been previously stated that interaction between participants is a key factor in focus groups and can provide very rich and useful data. Wilkinson[318] looked at women diagnosed with breast cancer and was interested to learn about their experience of treatments and surgery. It was found that the women could talk openly about very personal experiences and they even joked about common encounters, meaning that as well as gathering data the focus group was also a therapeutic exercise for the participants.

Stevens[319] carried out an additional piece of work that very much focused upon the interaction between individuals. Stevens looked at gathering information at the aggregate level. The focus group was used as a method of validating what was said in individual interviews conducted previously. Stevens states that not only can focus groups be used to gain in-depth insights but can be directed at segments of the population, which may have been under-served. A total of 13 lesbian women were asked to share their health experiences. It was found that during the discussions the women were able to talk freely about their experiences and could validate one another's concerns. As OE questions were asked the women had the freedom of directing the discussion and could talk about issues that were important to them; this is stated in the text as validating the data. Individual interviews also took place, demonstrating a triangulation of methods and enabled comparisons to be made between the different data collection methods. The group facilitator carried out the analysis and therefore may have been introduced observer bias, although a familiarity with the results and language used enabled themes and discrepancies to be dealt with. Again, the subject matter was considered to be acceptable to the respondents; this is evident because of the women volunteering

personal information and the free flowing nature of the discussions. This study was not intended to be generalisable to a wider population, rather to elicit responses from a specific group so as to incorporate their views on health topics that may have been missed during previous research. The women were all, however, very well educated and therefore selection bias may have influenced the results even though attempts were made to include women from different ethnic backgrounds.

In the study by Keller and colleagues,[320] 22 senior citizens were invited to attend focus groups to investigate their healthcare needs and beliefs. In addition, 16 22–40-year-olds were asked to express their opinions on the topic. The facilitator of the discussions used a set of nine OE questions defined before the sessions commenced. This is in contrast with the previous study[319] in that the issues to be discussed were predefined, rather than being led by the participants themselves. The sessions were all recorded and subsequently transcribed, although again it is unclear who carried out this task, and indeed if more than one researcher was involved. Follow-up calls were made to participants to give the opportunity for individuals to provide feedback and to validate what they had previously stated.

Smith and West[321] also sought the opinions of senior citizens. In all, 124 older people were asked about their experiences and expectations of health. All of the groups were facilitated and transcripts were analysed by two researchers who ensured that all relevant topics were covered in adequate detail. The focus group was found to be an acceptable method that raised some important issues in the provision of care to older people.

Bradley and colleagues[322] conducted a study where attitudes towards England's health strategy were explored. Here, 173 people took part in 24 focus groups. Again the groups were organised in a homogeneous manner, although the reasons for this are not stated. Reasons for exclusion of certain citizens were provided; that is, those with a history of HIV or AIDS or registered injecting drug users were not invited to participate as lay perspectives were sought. Including those people with a specialised knowledge may have swayed the results. In addition to this, unnecessary distress may have been caused for limited benefit to the study. Issues covered in the different groups included coronary heart disease and stroke, accidents, mental illness and sexual health. The validity of the analysis stage was established because two independent qualitative researchers carried it out. The cost of each 90-minute focus group is stated as being £365, taking into consideration recruitment, facilitating, recording, transcription and analysis. The generalisability of the findings is not mentioned but it is concluded that the method was inexpensive, that participants were able to grasp themes and that it was feasible and therefore acceptable. The authors express that this is a valuable method, which should be utilised when gauging citizens' opinions on healthcare.

The effect of discussion and deliberation on the public's views on priority setting was examined by Dolan and colleagues.[323] The aim of the research was to establish how much views had changed after deliberation. A total of 60 people were randomly selected to take part in the groups that consisted of five to seven people; the samples were stratified to include people from all ages and a balance of gender in each group. All participants were invited to attend two meetings, held 2 weeks apart, and were assessed as to whether their views had changed in the period of time between the beginning of the first group and the end of the second one. The two researchers who ensured that all relevant issues were covered also conducted all the groups. It was stated that to ensure external validation the sessions were tape-recorded and transcribed, although it is unclear who carried this out. It was concluded from the study that people's opinions on healthcare issues changed considerably after the opportunity for discussion and deliberation had been provided, and therefore valid results will be obtained after such discussion. The focus group method was compared with surveys in the text and it was concluded that these methods generate similar results; however, in this review the finding is that this is not the case and that the method used can have a significant effect on the results obtained.

Additionally, Ramirez and Shepperd[324] found that a questionnaire presented to participants was useful on two counts: first, to gather demographic information on those taking part and, secondly, to prompt people to think about the issues to be discussed, thereby allowing a period of considered thought. Also, informing people that the information gathered during the groups would be used to implement new health programmes was perceived to be advantageous when recruiting people. It is stated that people may be more willing to take part if they feel that if what they say will be used to improve healthcare in their own area. This particular group sought knowledge of and attitudes towards different risk factors associated with cancer

with a view to implement an intervention to promote healthier living in the local area.

The issue of analysing transcripts was dealt with by Weinberger and colleagues.[325] The study was concerned with whether or not the rating system of transcripts was consistent amongst researchers. A total of 101 ischaemic heart disease patients and 29 physicians participated in focus group discussions. Patients' experiences of care and the physician's decision-making processes concerning treatment procedures were addressed. The groups were recorded and transcribed before asking six health professionals to formulate a list of factors mentioned as being involved in decision-making for cardiovascular disease. Three raters were asked to categorise these factors as being either minor or major, to ascertain if the raters were consistent in their judgements. Each transcript was rated by at least one physician and one non-physician to ensure that both clinical and non-clinical perspectives were considered. It was concluded that consistency in raters' judgements was difficult to achieve, even with trained raters, and this therefore seriously undermined the reliability and validity of focus groups. It was suggested that it might be necessary to quantitatively analyse transcripts, although it was noted that this might be an unpopular suggestion to qualitative researchers.

In conclusion to this section, it can be seen that focus groups have been used widely in healthcare; a selection of examples has been shown here. These show that the focus group is a particularly useful tool in discovering people's opinions and attitudes towards their healthcare. The data provide explanations as to why people adopt certain perspectives towards their healthcare, and is therefore appropriate for exploring priority-setting issues.

Concept mapping

General methodological issues

Southern and colleagues[326] note that this a useful and acceptable way to gain group opinion and Trochim and Linton[327] note that researchers have no control over the final outcome, thus aiding objectivity. However, Trochim and Linton also state that it is a time-consuming and costly exercise but they also stressed that, with the development of new technologies, these costs can be reduced. They also stated that more work is need in this area to establish strengths and weaknesses of the process. Given that the technique involves feedback to respondents, there is the opportunity for respondent validation, therefore increasing the overall validity of the process.

Citizens' juries

General methodological issues

It has been argued that citizens' juries offer information, time, scrutiny, deliberation and independence.[331] A major strength is that informed deliberation can occur.[331] Lay people are given the opportunity to listen to, absorb and deliberate over information coming from several angles and from a range of professionals. People's rational deliberation is relied upon as opposed to a knee-jerk reaction.[581] It is felt that the deliberation process increases validity and reliability and that the method is particularly suited to difficult and complex questions.[331] Some jurors have expressed, however, that it would be beneficial for the time available for deliberation to be extended and perhaps a break would enable them to process the vast amount of information received.[333] In addition, strengths are "clear aims and role for jurors, a mechanism for implementing jury recommendations, and the incorporation of features such as information provision, time to question witnesses and deliberate, and a degree of independence and authority". These features indicate that the method is an acceptable one to the participants.

Citizens' juries have, however, recently been criticised as being "morally and democratically irrelevant".[581] Price argued that individual issues or case studies cannot be generalised to other issues and that it is incorrect to assume that principles drawn up for one situation can then be universally applied.[581] In addition, Price believes that participants are asked to perform an impossible task. They are, on one hand, expected to be neutral and impartial but, on the other hand, are tax-paying members of the community "committed to family, friends, his or her city, or even the NHS". Price continues, "Citizens are at once disinterested and committed. They contemplate broader issues and yet are motivated by relatively narrow ones."

Citizens' juries can be a very time-consuming and therefore expensive process, especially since it takes a very limited number of people's views into account. The planning of the exercise is a very important component and therefore adequate time needs to be set aside.[331] A 4-day jury can cost as much as £25,000 (in 1997); this included expenses such as recruitment, incentives for the jurors, venue, development of an agenda, recruitment of witnesses, moderators and

publication and dissemination costs. Clearly, this is a drawback; however, if carried out in a rigorous manner, the cost can be justified because it obtains a more informed opinion of the public than a quantitative survey.[328]

A further potential for bias or lack of objectivity in citizens' juries is for the witnesses to manipulate the jury. In one study a person with no knowledge of the subject matter was selected as moderator to ensure that the deliberations were as objective as possible.[311] A further disadvantage is that only some issues can be covered. Citizens' juries have been shown to be good for choosing between two options and solving dilemmas, but not where there is a need to develop detailed plans or discuss hypothetical situations.[331,332] They can deal only with problems that can be addressed by a decision and a set of recommendations at a particular point in time[329] and are unable to look at problems that need frequent review.[582] Jurors do not set the agenda or are not able to bring in additional issues as they see fit, which could potentially alter the outcome.[329] Currently, citizens' juries are not used as a forum for joint discussion with the priority decision-makers. This means that the process does not allow for members of the jury to meet and discuss relevant issues with the policy-makers themselves.[329]

Examples of the application of the technique

Citizens' juries are a relatively new way of finding out the public's opinion on priority setting. A series of five pilot juries was carried out in 1996. One took place in Cambridge and Huntington; two in Kensington, Chelsea and Westminster (KCW); and one each in Walsall and Luton. Three additional juries sponsored by The King's Fund followed these up in Sunderland; East Sussex, Brighton and Hove; and Buckinghamshire. Each of these juries discussed different issues concerning healthcare policy and healthcare rationing.[328] The first group to use it as a research tool was the Institute for Public Policy Research (IPPR) in 1996 with Cambridge and Huntington Health Authority.[328,331,332]

The citizens' jury in Cambridge and Huntington Health Authority aimed to (1) gather information on how the jury members answered questions relating to priority-setting issues; and (2) evaluate the process of citizens' juries as a method of eliciting public opinion. The 16 jurors were recruited using stratified random sampling to get a representative sample.[332] They were asked to address questions about whether the public should be involved in priority setting, the criteria that should be used to make decisions about healthcare and who should set these priorities.[331] Over a period of 4 days jurors heard information from expert witnesses and were given the opportunity to question and cross-examine them and could deliberate over their decisions. By asking the jurors to fill out questionnaires both before and after the process, information was obtained about the successes and failures of the technique.[332] One problem noted here is that one or two highly knowledgeable and/or very vocal members of the group can limit the opportunity for everybody to express their opinion, a problem to be found in a number of qualitative group-based methods such as focus groups and public meetings. On this occasion this difficulty was overcome by splitting the jury into two smaller groups to give the quieter members more of a chance to speak without feeling intimidated.

Kensington, Chelsea and Westminster Health Authority held two pilot juries to debate the issue of how the quality of life of mental health patients could be improved. The Walsall jury investigated how £600,000 could be best spent to improve palliative care. Finally, the Luton jury were asked "how should citizens pay for health services in the future?"[331] In The King's Fund projects, the Sunderland jury was asked about whether local people would accept certain treatments from health professionals other than their GP. East Sussex, Brighton and Hove were asked to consider where women should be treated who have been diagnosed with gynaecological cancer. In Buckinghamshire, jurors were asked whether the health authority should fund treatment for back pain.[333]

These pilot studies were evaluated by Sang (cited in Davies and colleagues[333]). A number of comments were made about the effectiveness and practicalities of the process. It was ultimately concluded that a group of citizens were able to work well together as a team and could appreciate the complexities of a health policy issue. The health authorities viewed the jurors' recommendations positively and the jurors viewed the experience positively. However, three complications were detailed. First, there is the concern that local people may have difficulty engaging with the major policy issues facing the NHS; secondly, the role of health authorities in making these types of decisions is not yet clear; and finally, it is felt that the more the issue of involving citizens in decision-making is investigated the more complex it becomes, and the more questions need

answering. Coote and Lenaghan[331] concluded that provided the question being asked is "prepared in an appropriate and manageable form" then the participants are willing and able to make sensible judgements about complex issues. It was realised from the Luton jury that presenting four options of how to fund the NHS was too challenging and it was suggested that three or two would be more manageable. They also felt that the jurors acted on behalf of community interests and not personal ones and that jurors' confidence grew throughout the proceedings.

The Welsh Institute for Health and Social Care adopted the citizens' jury approach to investigate new genetic technologies.[334] The jury process was conducted in a similar manner to those described above but had three major differences. First, a multinational pharmaceutical company sponsored the jury rather than a policy decision-making body. This meant that the recommendations made by the jury did not have to be adopted, and they did not have to offer explanations about why they would not do so. Secondly, the nature of the question being addressed meant that an understanding of material of a scientific nature was required. The jurors were provided with educational material but this was rarely utilised, thereby questioning whether the participants fully understood the issues. Thirdly, those selected to take part were to represent Wales and not a small local community, but doubts were expressed as to whether this type of tool was appropriate for this level of policy recommendation. As in some of the previous juries the issue of representation was a concern for both the organisers and the jurors themselves. In this instance only one member had continued their education after reaching the age of 18, there were no participants from ethnic minorities or from a young age group and few of the jurors were in full-time employment. This not only puts a question mark against the representativeness of the jury but also its generalisability. The moderator chosen was highly experienced at working with small groups such as this. He attempted to ensure that all members took part in discussion, that time schedules were adhered to and that outcomes of discussion were recorded. It is not made clear if he had specialised knowledge of the subject matter and therefore the extent of his objectivity cannot be commented upon. It would be improved, however, by including a second moderator. This jury highlighted a number of other issues relating to the potential for bias. There was a concern that some of the witnesses were introduced using their scientific titles but that other non-scientific members, who would be of equal importance, were referred to by name only. One lay witness felt that this undermined her status and felt "down-graded". Additionally, much consideration should be given to the ordering of witnesses. Here one medical witness appeared both at the very beginning and at the very end of the process, which could have had implications for validity and reliability. The participants and authors viewed it as a positive experience but felt that "the primary justification for using the citizens' jury approach was educational rather than participative". They underlined the notion that some people feel uncomfortable about being involved in these types of decisions but yet felt that it was useful as an educational tool. The authors concluded that "as yet there is little evidence that the recommendations are more than a one-off snapshot view produced by a small, but enthusiastic, group of inhabitants of the principality".

Since the technique was developed it has become increasingly popular as a tool for aiding decision-making. The Portsmouth jury, held in 1997, addressed the question of what are the most important criteria for setting spending priorities.[328] To help the jury to test their criteria they were given descriptions of four treatments and asked to allocate a set budget of £80,000 (as such this can be seen as a variation of the allocation of points method within a citizens' jury). The Nottingham jury was presented with the question: Should non-clinical factors be taken into account when prioritising NHS resources?[328] Lewisham held a jury asking the question: What can be done to reduce harm to the community and individuals from drugs?[335]

Consensus panels

General methodological issues

As this method is similar to the citizens' jury, the same methodological issues of discussion leaders, selection methods and deliberation need to be considered. An advantage over citizens' juries would be the lower cost of the exercise, which may then make it a more practical and acceptable method to use.

Examples of the application of the technique

A UK study conducted by Coulter and colleagues[340] investigated the effect that panel composition had on the ratings for use of spinal manipulations for lower back pain – a methodological issue with implications for validity, reliability and generalisability. One panel consisted of a multidisciplinary team and the second panel consisted only of

chiropractic physicians. It was concluded that a multidisciplinary team was more likely to rate a treatment inappropriate than a team of professionals who actually used the treatment or technique. The three other studies did not detail how panels were selected or highlight methodological issues. A study that did detail these issues was carried out by Stronks and colleagues.[336] They conducted a study in which the views of patients, the general public, GPs, specialists and health insurers were incorporated in a consultation to achieve a one-third cut in the healthcare expenditure. The panel members were given information on ten scenarios of services that could be dropped or not funded. The members were asked to discuss these issues and decide on areas where cuts should be made. Panels were not selected to be representative of the population, but chosen to represent five distinct groups. The authors recognised that this would affect validity but highlighted that it was not their aim to be representative but rather to gain insight into the arguments and concerns of different groups of people when considering priority issues. A facilitator was appointed. It is not stated how this person was selected, and his or her personal views could have influenced outcomes, thereby introducing bias and questioning validity and objectivity.[336]

According to the authors this technique has limited value because people had a lack of knowledge relating to the issues they were asked to discuss and were not given enough time to absorb information and form well thought out opinions. However, this study was carried out using particularly controversial subject matter. The conclusion drawn from the study was that "it is not clear that including all the different actors in the decision-making process of prioritisation of health services will lead to more equitable or broadly supported outcomes or to better health to the population". It was established in this study that healthcare professionals were more aware of the importance of equal access to health than were other groups, in turn suggesting that the public and patients are not best equipped to make these types of decisions when using the consensus panel method.

Public meetings

General methodological issues

Public meetings have been shown to be an inexpensive and relatively quick method of obtaining information but can be severely criticised in that only those with a particular interest or who are very vocal about it will attend,[583] and those attending will differ from the general population in demographic characteristics.[341] Meetings are often poorly attended and there is the potential for them to be dominated by specific interest groups.[328] Additionally, they have been criticised for addressing issues that are not of the greatest concern to the public.[583]

Public meetings have often been used as a method of making people aware of potential closures of health facilities, therefore generating a feeling of hostility to them.[583] The more cynical may argue that perhaps information-giving is the main aim of public meetings under the pretence of listening to the public or that they are used to legitimise decisions that have already been made.[584]

The volunteer bias questions the validity, reliability and generalisability of the method as the sole means of eliciting public opinion. However, it is recognised that they may still be useful if used alongside other methods to gauge opinion.[583]

For public meetings to be as reliable and valid as possible it is recommended that they are well publicised, that there is easy access to all members of the public and that all those in attendance are asked for their opinion. Jessop[585] felt that for NHS meetings these first two criteria are usually met but the third is neglected, and that gaining a mere impression of general mood is not acceptable or reliable.

Examples of the application of the technique

Gundry and Heberlein[341] investigated hypotheses suggesting that these types of meetings are unrepresentative. They found that although it was true that people attending did differ in their demographic details the overall information gained was the same as that from a survey of the general population. It was also found that those with extreme opinions might well attend the meetings but that the opposite can also occur, in that those who attend can have stronger agreement with the general public than those not attending. The authors argued that "public meetings may be a useful and valid tool for capturing a reasonably accurate picture of public opinion on a variety of issues". It is, however, recognised that representativeness cannot be assured unless a simultaneous survey is conducted alongside the meeting, which would then compromise the major benefits of low cost and ease of implementation that public meetings can offer.

In the Broadbent study,[342] issues surrounding advertisement and access to public meetings in

her local area were explored. She found a lot of variation among the 12 trust boards she contacted. Seven were helpful and all the information requested was obtained from one telephone call; for the other five, however, considerable difficulty was experienced in finding out when and where the next meeting would be held. She also attended the meetings and found that a number of them were, first, difficult to find and, secondly, provided no opportunity for public participation. It was her recommendation that each and every meeting be adequately advertised, directions of how to get to the location provided, and board members be introduced; those attending should receive a written agenda and time should be allowed for comments and questions.

A number of public meetings were set up and reviewed in a study carried out by Gott and Warren.[343] The community of the North Staffordshire district formulated a Neighbourhood Health Forum, which was designed to increase local participation in health issues and used public meetings as a way to find out the feelings of the community. To make people aware of the meetings and to increase attendance, selected members of existing organisations, mainly voluntary groups, were personally invited. The meetings were also advertised, inviting anyone interested in being involved in local healthcare decisions to attend. They were held in buildings frequently visited by the public in an attempt to encourage participation. It was found that there were a number of professionals in attendance who were not directly working in the health sector. This suggested that attempts made to involve a diverse range of community members were successful, although no mention was made of the more marginal groups such as unemployed people, ethnic minorities or people with disabilities.

This study showed that there is value in conducting these types of meetings and that extensive steps were taken to maximise attendance and thereby generalisability. This is further illustrated with the suggestion that information sheets, newsletters and personal invitations be distributed to households in the district. There is, however, still a dearth of studies that have evaluated methods of recording and analysing the content of most methods of public consultation.

Nominal group technique

General methodological issues

Similar considerations concerning the facilitator and participant selection are required for the nominal group technique as for the methods described above. There is a high potential for selection bias[348] and the composition of those people taking part can have an effect on the final outcomes. There is also the potential for false consensus to be obtained, especially in situations where there is diversity of opinion on priorities.[347] Redman and colleagues,[349] in their study looking at priorities in breast cancer service provision in Australia, note that the method appears to have strong validity because there was agreement about priorities amongst the different workshops. Additionally, validity was demonstrated when results were presented to government agencies, clinical specialists and consumers who agreed with the priorities identified. Redman and colleagues also note that the method is acceptable to participants who were willing to participate and were satisfied with the process.

Discussion and conclusion

The systematic review identified a number of qualitative techniques that have been, or could be, used for eliciting public preferences. Techniques were identified as individual or group-based.

One-to-one interviews are an extremely useful tool for obtaining information from individuals about their perspectives on health and healthcare. The advantages are that: the interviewer is present to clarify issues and provide verbal and visual clues; in many cases more personal information can be discussed; each participant is provided with adequate time to fully express himself or herself. The major problems with this technique are that only small samples can be examined because of the detailed information required and that the method is very open to interviewer bias in respect of the topics discussed. Ways of minimising these problems are suggested in the literature and have been highlighted here. A number of new methods outlined here, namely telephone, email and dyadic interview techniques, have not yet been used in priority setting. Given their potential benefits, they worth investigating.

The Delphi technique is a flexible method that enables the opinions and judgements of experts to be collected. Respondent bias is minimised owing to the anonymity of the technique, which eliminates peer group pressure, although there may be a high level of drop-outs because it can be a three or four round process and therefore only the most committed may remain. Issues of reliability, validity and generalisability can be affected by the selection of the panel members. On the issue of objectivity, the method is liable

to researcher bias, and the minimisation of this is the responsibility of the individual researcher as is the case in other qualitative methods. An advantage to this method is that it is a cost-effective way of gaining a large scope of opinion from different groups of people situated in various locations. It is recommended in the literature that it may be necessary to spend a little more and engage in an initial investment of time in this kind of meeting in order to increase the chances of participation. This in turn will increase the validity and generalisability of a study and fulfil the criterion of acceptability to respondents. Although most of the studies identified involved asking groups of health professionals, we conclude that the Delphi technique is sufficiently reliable and valid to be used more widely to gain information from patients and from the general public.

With regard to focus groups, there is ambiguity concerning the ways in which focus groups should be conducted and whether this is a cost-effective way of acquiring opinion. The more recent literature suggests that each study should be considered individually and that a set of rules can work against this method because it can be restrictive. There is, however, consensus concerning their pitfalls. They do not use large samples, and are therefore not generalisable to wider populations, although this is not always the aim. Sometimes focus groups seek the opinions of specific sections of the community to get a feel for the types of issues they consider to be important and there is no desire to project these ideas to large populations. They are susceptible to selection bias, in that only those with very strong feelings on an issue will participate, and in addition to that the most confident people will be heard whereas the quieter members of the group may feel intimidated. Group composition can affect the data gathered. Careful consideration must be given to whether it would be an advantage for participants to know one another, or to be complete strangers; if people feel uncomfortable then they may withhold information. Biases can also occur through the facilitator of the discussions. Objectivity and skill of the researcher is essential in order to maximise reliability. It has been shown that some sensitive topics are acceptable for group discussion; however, participants must feel comfortable with talking about these types of issues with both the other members of the group and the facilitator. In this way previously untapped information can be uncovered. If participants feel uneasy about disclosing personal information, they may withhold it and therefore affect the validity of the study.

Citizens' juries are a relatively new technique for eliciting preferences. Their major strength is that people have the chance to deliberate over some very complicated and involved topics. Participants are provided with information from a number of professional and lay people and can discuss with other jury members their beliefs and personal conclusions. For this method to succeed each jury must be carefully conducted. It is a valuable tool if moderators are carefully selected and are highly trained and experienced, and if a representative cross-section of the population is considered. If these basic recommendations are not adhered to, then the jury may fail to reach conclusions on these difficult and complex issues. It is also very important to remember the criticisms presented by Price, namely that citizens' juries are morally and democratically irrelevant because they are not representative.

From the evidence presented it would appear that consensus panels are not an appropriate method of involving the public in priority setting. However, it may be necessary to conduct further studies using this tool before it can be dismissed. It is less costly than a citizens' jury and could be a valuable technique if panel members are carefully selected and given adequate time for deliberation. A crucial issue not yet addressed, however, is whether the questions being asked are those that lay people would ask.

Public meetings are commonly used and are a quick and cheap way of gaining opinion, but the disadvantages appear to far outweigh these advantages. They can be unrepresentative in that they only capture the opinions of those with a special interest and those willing to speak out. Generally, the public has become sceptical about the motives behind these meetings because they have been used to inform people of a decision that has already been made. Additionally, it has been noted that they can be inaccessible to the public because of failure to advertise and poor organisation. Despite this, it is recommended here that they should not be discounted until there has been further evaluation of what they add when used with caution and alongside other methods.

Complaints procedures are noted but these will only consider particularly negative view points, and have limited value in assessing public opinion. Further work is needed to establish the appropriateness of nominal group techniques, concept mapping and case study analysis.

Chapter 7

The importance of public views in priority setting: the perspective of the policy-maker

So far this review has identified potentially useful techniques for eliciting public preferences in the delivery of healthcare. Such an exercise is only useful if those views are taken into account, or at least deemed important, in the decision-making process. This chapter reports the methods and results of a pilot study set up to address this question. Although both qualitative and quantitative techniques have been explored in this report, this chapter adopts a mainly quantitative approach to establishing the weight attached to public preferences *vis-à-vis* other criteria used to set priorities. The next section presents the methodology of the study. The study was divided into two stages: a questionnaire-based survey including **choice-based CA**[*] and an **allocation of points exercise** to establish the weights of the different criteria[†], and a telephone interview/follow-up questionnaire to further explore views concerning involving consumers in decision-making, as well as views concerning the different techniques for eliciting weights. The results are then presented and discussed.

Methods

Questionnaire-based survey

The literature review described in chapter 2 led to the identification of literature on priority setting. This literature was read with a view to identifying criteria, other than consumer views, which had been used in priority setting. When reading this literature consideration was also given to frameworks for priority setting. The results from this review are shown in appendix 5. Following this, discussions took place between one of the researchers (AB) and health professionals involved in priority setting to identify a set of criteria used for priority setting. The criteria included in the study are presented in *Table 1*.

In this chapter the focus is on the community values and priorities criterion, and to what extent evidence citing the views of the community is incorporated into priority-setting decisions. Three additional criteria – the severity of the disease/condition, how many people it affects, and whether national/local priorities are fulfilled – were incorporated into the questionnaire by including them in the scenarios presented. Two scenarios were presented: the first involved a health service for gynaecological cancer and the second a health service for asthma. This design was based on the *a priori* hypothesis that different preferences may exist under the different conditions. A copy of the questionnaire is given in appendix 6.

Choice-based CA

When using choice-based CA, levels must be assigned to the criteria. These were decided with reference to the literature review and discussions with those involved in priority setting (see *Table 1*).

The criteria and levels defined in *Table 1* gave rise to 9375 ($5^5 \times 3^1$) possible combinations of healthcare proposals. Experimental design tables[586] identified an orthogonal design of 25 proposals, which were paired into 13 choices. Given that two different scenarios were presented,

[*] Also known in the literature as CA, stated preference and stated preference discrete choice modelling, and discrete choice experiments.

[†] These two approaches make different assumptions about how individuals form their preferences. In the first, it is assumed that whilst respondents can provide an overall evaluation of whatever is being valued, they do not have the ability to link the contribution of the individual components (or weights of these components) to the overall evaluation. Assuming this approach, the researcher must infer the weights that are implicit in the respondents' evaluations. In contrast, the second approach assumes that respondents know the individual weights they assign to the criteria or characteristics of the good being valued. Whilst there has been a large amount of psychological work comparing these two approaches to decision-making,[562] no such work exists in healthcare. This study compared the two approaches.

TABLE 1 *Criteria defined with their corresponding levels*

Criteria	Levels for discrete choice experiment
Potential health gain (HEALT)	• Life-saving now • Sustained improvement now • Sustained improvement later • Temporary improvement now • Temporary improvement later
Evidence of clinical effectiveness (CLIN)	• Empirical evidence (MA, RCT, descriptive) • Expert opinion • None
Budgetary impact (BUD)	• Big save • Small save • None • Small expense • Big expense
Equity of access and health status inequalities (EQUIT)	• Big reduction in inequality • Small reduction in inequality • Remains the same • Small increase in inequality • Big increase in inequality
Quality of service (QUAL)[a]	• Two or more direct 'hits' • One direct 'hit' • Two or more partial 'hits' • One partial 'hit' • No 'hits'
Community values and priorities (COM)	• Robust evidence 'support' • Weak evidence 'support' • Robust evidence 'indifferent' • Weak evidence 'object' • Robust evidence 'object'

[a] *See the questionnaire in appendix 6 for definition of a 'hit' in terms of the quality of the service*

MA, meta-analysis

this resulted in 13 choices for each scenario. These 26 scenarios formed the basis of two separate questionnaires (Type 1 and Type 2) consisting of 13 discrete choices. The Type 1 questionnaire included six choices from the first scenario and seven from the second, whereas the Type 2 questionnaire comprised the remaining seven choices from the first scenario and the remaining six from the second scenario.

Two dominant choices were included in each questionnaire to assess the internal consistency of responses (definition in chapter 4). Respondents were dropped from the analysis if they 'failed' both tests on the basis that if they failed one test this may be because of random error whereas if they failed both tests this is more likely to represent a lack of understanding of the process.

From the choice-based responses, the following benefit equation was estimated:

$$\Delta V = \alpha_1 HEALT + \alpha_2 CLIN + \alpha_3 BUD + \alpha_4 EQUIT + \alpha_5 QUAL + \alpha_6 COM + e$$

where ΔV is the change in benefit (or utility) in moving from proposal A to proposal B, and the independent variables are the differences in the levels of the criteria, as defined in *Table 1*. The α's are the coefficients of the model to be estimated, indicating the marginal importance of a unit change in the given criterion, and e is the unobservable error term[‡].

The above equation was used to estimate the weights of the different criteria: α_6 is the weight associated with the public's views criterion and α_n the weights of all other criteria ($n = 1, 2, 3, 4, 5$). These weights indicate the marginal change in overall utility, V, resulting from a marginal change in this given criterion. This marginal change is obviously dependent on the unit of measurement. So, for example, whilst the coefficient on potential health gain indicates the marginal change in benefit of moving from say 'temporary improvement now' to 'sustained improvement later', the coefficient on 'community values and priorities' indicates the marginal change in benefit in moving from say 'weak evidence "object"' to 'robust evidence "indifferent"'. The Wald statistic was used to test whether the criteria weights were significantly different using the following null and alternative hypotheses:

$$H_0: \alpha_6 - \alpha_n = 0$$

$$H_A: \alpha_6 - \alpha_n \neq 0$$

A general-to-specific approach was adopted. The general model includes all the criteria, whether or not they are significant. The most significant variable is then dropped and the model re-run, repeating the same process until a model with only significant attributes at the 5% level remains. This latter model is referred to as the specific model. The Chow-type likelihood ratio test was

[‡] Given the nature of the experimental design employed, interaction terms could not be considered.

used to investigate whether the weights for the different criteria differed according to the scenarios presented to respondents (gynaecological cancer or asthma). If there was no difference then the data sets could be merged and jointly analysed.

Allocation of points

In this study, points were used to define the budget. Respondents were asked to allocate a total of 100 points (deemed a manageable amount[563]) between the criteria identified. In order to encourage honest preference revelation and to explicitly incorporate strength of preference into this exercise, respondents were reminded that the points they allocated to the different criteria should total 100 and represent their strength of preference. An example of this type of question is illustrated in appendix 6. In the analysis the mean number of points allocated to each of the criteria across all individuals represented the weight and hence the importance of that attribute. Analysis using the median was also included and the results compared. Paired *t* tests were used to test for statistical differences across the two scenarios. If there was no difference, the data sets could be merged and jointly analysed. Using similar hypotheses to those illustrated in the discrete choices section, paired *t* tests were also used to test whether the weights placed on the criteria were significantly different compared with the community values criterion (at the 5% level). The Friedman test was used to test whether there were significant differences between the way the criteria were ranked across respondents.

Sample and setting

The questionnaire was sent to 109 health policymakers and healthcare professionals in Scotland: 46 were employed at the health board level; 24 at healthcare trust level; and 39 comprised a convenience sample of University of Aberdeen Health Economic Correspondence Course students (the typical background of such students includes managers and providers of healthcare).

Telephone interviews and follow-up questionnaire

Respondents to the questionnaire were asked if they would be willing to be contacted to discuss their responses. Those who agreed were initially sent a letter reminding them of the study with a copy of their completed questionnaire. They were informed that they would be contacted by telephone. A structured telephone interview schedule was used (appendix 7). The convenient sample of correspondence course students completed the interview schedule as a follow-up questionnaire. The aim of this further questionnaire was to directly question respondents on the importance of public views in the priority-setting process and to assess their views on the two techniques used to elicit criteria weights. Respondents were also asked to rank the criteria so that their responses could be compared with the results from the questionnaire-based survey.

Results

Questionnaire-based survey

Of the 109 questionnaires sent out, 57 were returned in the required 6 weeks (with no reminders owing to the timescale). Of these, five were returned uncompleted. The remaining 52 respondents all passed the dominant tests and were included in further analysis, giving a usable response rate of 48%. Bias due to non-response may be a problem here given the relatively low response rate. This could be investigated in a larger study by comparing the characteristics of respondents and non-respondents and analysing whether the samples were statistically different from one another. In this pilot, specific characteristics were not available for non-responders.

Table 2 shows the professional background of the respondents.

Choice-based CA

The Chow-type likelihood ratio test did not reject the null hypothesis of homogeneity; that

TABLE 2 *Profession of respondents to the questionnaire*

Profession	Frequency
Health professional employed at healthcare trust level (e.g. associate medical director, clinical GP coordinator, head of service)	13 (25%)
Health professional employed at health board level (e.g. PHC, HE)	19 (37%)
Health Economics Correspondence Course students	20 (38%)

PHC, public health consultant; HE, health economist

is, the hypothesis that the coefficient results are the same irrespective of the scenario under which they were presented. The results gained under the two different scenarios could therefore be merged and presented as one. The results from the regression equation are shown in *Table 3*.

The results from the specific approach show that, with the exception of quality of service, all criteria had a significant effect on the choice of the health proposal. All criteria had a positive effect, indicating that the better the level of the attribute, the more likely the respondent would choose that healthcare proposal. The weights of the criteria are given by the coefficients. A higher coefficient indicates that a unit change in that criterion had a higher effect on benefit. A unit change in evidence of clinical effectiveness was considered to be the most important of the criteria, followed by budgetary impact, community values and then potential health gain. However, there was no significant difference between the importance of community values and the other four criteria.

Allocation of points

Paired *t* tests indicated no difference in the allocation of points according to the two different scenarios. The results presented in *Table 4* are therefore for the combined data set.

Different marginal weights to those found using choice-based CA were obtained, with potential health gain and evidence of clinical effectiveness carrying the highest weights. Community values were the least important criterion. Paired sample *t* tests revealed significant differences (at the 5% level) between each of the criteria when compared with community values. The Friedman test indicated that there were no significant differences between criteria ranking across respondents; that is, health gain was ranked first (with the highest number of points) and community values last (with the lowest number of points) a significant number of times. Using median values gave similar results to those found using the mean values with the exception that both community values and quality of service were deemed equally to be the least important criteria.

Telephone interviews and follow-up questionnaire

From the returned questionnaires, 15 respondents were willing to be contacted at a later date to discuss their responses in a telephone interview and in addition 11 Health Economics Correspondence Course students completed a follow-up

TABLE 3 *Results from choice-based CA*

	General	**Specific**
Criteria	Coefficient (weight)	
Potential health gain (α_1)	0.3992**	0.3067**
Evidence of clinical effectiveness (α_2)	0.4773**	0.4756**
Budgetary impact (α_3)	0.4188**	0.3265**
Equity of access and health status inequalities (α_4)	0.3807**	0.2877**
Quality of service (α_5)	0.1206	–
Community values and priorities (α_6)	0.4082**	0.3202**
Number of observations	671	671
Log-likelihood function	–294.9988	–297.3248
ACI[a]	0.897	0.901
Restriction/null hypothesis	**Chi-squared (*p*-value)**	
H_0: $\alpha_1 - \alpha_6 = 0$	0.14 (0.71)	
H_0: $\alpha_2 - \alpha_6 = 0$	3.51 (0.06)	
H_0: $\alpha_3 - \alpha_6 = 0$	0.04 (0.85)	
H_0: $\alpha_4 - \alpha_6 = 0$	0.42 (0.51)	

[a] *Akaike's information criterion (AIC) is a measure of goodness-of-fit. It is estimated as:* $AIC = -2ln L(M_\beta) + 2P/N$, *where* $L(M_\beta)$ = *log-likelihood of the model,* P = *the number of parameters in the model and* N = *the number of observations*

** $p < 0.01$

TABLE 4 *Results from allocation of points exercise*

Criteria	Minimum number of points	Maximum number of points	Mean number of points	Relative importance	Friedman test mean rank	Median number of points	Relative importance
Potential health gain	15.00	65.00	28.5	1	5.44	27.00	1
Evidence of clinical effectiveness	5.00	60.00	23.1	2	4.58	21.75	2
Budgetary impact	0.00	30.00	15.7	3	3.52	16.83	3
Equity of access and health status inequalities	0.00	30.00	13.3	4	3.04	15.00	4
Quality of service	0.00	20.00	11.0	5	2.53	10.00	5
Community values and priorities	0.00	20.00	8.5	6	1.89	10.00	5

Paired sample *t* tests	***t* test**	***p*-value**
Potential health gain – community values	11.063	0.000
Evidence of clinical effectiveness – community values	8.474	0.000
Budgetary impact – community values	6.032	0.000
Equity of access – community values	5.373	0.000
Quality of service – community values	2.901	0.005

questionnaire of the same structure. However, given time constraints, six people were contacted by telephone: two were healthcare professionals at trust level and four were healthcare professionals at board level. All 11 questionnaires from the students were used. Despite the small numbers involved, a number of interesting findings emerged from the follow-up interviews and the questionnaires (*Table 5*). Foremost, respondents felt that the important criteria in priority setting had been covered in the first questionnaire.

Regarding the choice-based CA approach to eliciting weights, 11 of the 17 respondents thought that this was a realistic or very realistic approach. Ten respondents found such choices very difficult or difficult. However, in retrospect, the term 'ease of making choices' does not distinguish between whether the questionnaire was difficult to understand or whether the questions were difficult to answer (i.e. whether the answers required careful consideration). The two are quite different. Although we may be looking for ease of comprehension, when we are considering a difficult subject such as priority setting, we would not expect the choices to be easy. This conclusion is supported by the fact that studies applying the technique to the general public and patients have reported respondents having little difficulty completing the choices.[191,194,199,201,202,203,396] Choice-based CA assumes that individuals consider all the criteria in the study, and that they are willing to trade between these. However, nine of the 17 respondents indicated that they focused in on some of the criteria. This suggests that future work should explore in more detail the decision heuristics that respondents employ when completing choice-based experiments.

Six of the 17 respondents thought that the allocation of points exercise was difficult or very difficult. Again, this may reflect the nature of the questions being asked. In the allocation of points task, individuals were encouraged to think about their strength of preference. However, the results suggest that this may not have been the case, and that the resulting weights may be ordinal.

Despite community values and priorities being considered the least important criterion in the ranking exercise (*Table 6*), 15 of the 17 respondents thought that the community had a role to play in priority setting and 12 indicated that this role was important or very important.

A summary of rankings obtained from the results of the different techniques is shown in *Table 7*.

TABLE 5 *Responses to the telephone interview/follow-up questionnaire*

Question	Responses			
Were any criteria irrelevant?	**Yes** 1	**No** 13		
Discrete choices				
	Very hypothetical	**Hypothetical**	**Realistic**	**Very realistic**
Realism of choices posed?	1	4	6	5
	Very difficult	**Difficult**	**Easy**	**Very easy**
Ease of making choices?	1	9	4	2
	Consider all the criteria		**Focus on some criteria**	
Decision heuristics?	8		9	
The allocation of points				
	Very difficult	**Difficult**	**Easy**	**Very easy**
How easy did you find allocating the points?	0	6	8	2
	Yes	**No**		
Reflection of strength of preference?	12	5		
Comparing the methods				
	The paired choice exercises		**The allocation of points exercises**	**Neither**
Which of the two methods did you find easiest?	5		6	4
Which of the two methods did you find quickest?	3		7	5
Priority setting and the community				
	Yes	**No**		
Do you think that the community has a role to play in priority setting?	15	2		
	Of no importance	**Of little importance**	**Important**	**Very important**
How would you rate community views when deciding whether or not to implement a proposal?	0	4	4	8

TABLE 6 *Criteria ranked in order of importance by telephone interview/follow-up questionnaire*

Criteria	Mean ranking	Rank
Potential health gain	1.1	1
Evidence of clinical effectiveness	2.8	2
Equity of access and health status inequalities	3.4	3
Quality of service	3.7	4
Budgetary impact	4.8	5
Community values and priorities	5.3	6

TABLE 7 *Summary of the rankings obtained from the results of the different techniques*[a]

Criteria	Choice-based CA	Allocation of points exercise	Direct ranking exercise[b]
Potential health gain	4	1	1
Evidence of clinical effectiveness	1	2	2
Budgetary impact	2	3	5
Equity of access and health status inequalities	5	4	3
Quality of service	–	5	4
Community values and priorities	3	6	6

[a] *Whilst the rankings for the choice-based CA represent the impact of a marginal change in the criteria, the allocation of points and direct ranking results are for the criterion generally*

[b] *Telephone interview/follow-up questionnaire*

Discussion and conclusions

A choice-based CA approach was used to estimate weights indirectly, and the allocation of points approach was used to estimate weights directly. Both approaches found the views of the public to be important in the priority-setting exercise, although the relative rankings differed across the two techniques. The results from the allocation of points exercise were similar to the direct ranking exercise carried out in the telephone interview/follow-up questionnaire. Fundamentally, the conclusion to be drawn from these results is that the weights assigned to the various criteria appear to differ depending on the elicitation method used. Following from this, the context in which the technique is to be used should be taken into account.

This may be an obvious, necessary conclusion for some. Cookson points out that a number of psychologists and behavioural theorists have argued that preferences are constructed in response to stimuli rather than revealed or discovered.[587] This implies that if preferences are constructed in different ways (given the different methods), different preferences will be elicited, implying predictably different results. In this case, the different assumptions made regarding the CA approach, the allocation of points method and the direct rankings may be reflected in the results. Expanding this further, Clark notes that eliciting precise results from the budget pie technique (termed allocation of points here) is dependent on the fulfilment of some basic assumptions.[277] These include issues surrounding the use of a monetary budget, the elicitation of honest preferences and the revelation of intensity of preference.

The problems encountered with the use of money in the allocation of points method were avoided by using a points system. The main focus of discussion is therefore honest preference elicitation broadly and, intertwined with this, the elicitation of the intensity of preferences. First, honest preference elicitation is not a problem specific to the allocation of points exercise; however, within this method there is room for strategic games to take place. This is likely to happen if respondents to the exercise are highly informed about the subject area and the subject matter is of interest and highly important.[277] In this study, it is assumed that respondents will report honest preferences. The results from the exercise reveal little variation between the mean and median criteria values indicating, to some extent, that there were few outlying strategic values and most people reported honest preferences.[277] This is not particularly surprising because it is not clear, in this case, what benefit would ensue from strategic behaviour. However, as a method whereby the weights are elicited directly, it is susceptible to such behaviour. In contrast, it may be argued that the implicit elicitation methods of the choice-based CA methodology would incite low strategic behaviour. However, even when a robust design is used there is still a concern by some authors who believe that rational individuals will never truly reveal their preferences.[282]

Secondly, it is debatable whether the allocation of points methodology can successfully incorporate or take account of the intensity of respondents preferences. The telephone interview/follow-up questionnaire revealed that a number of respondents (5/17) did not allocate points in a way that reflected their strength of preference. Elsewhere, Hoinville and Courtenay carried out an exercise

where the number of points they used was successively reduced and they revealed that respondents' allocations to the various options were not reduced proportionately.[564] Although there are other issues that cloud the reasoning on why this should happen, such as repeating the experiment and the optimal number of points that should be used, one interpretation is that respondents' strengths of preference are not indicated using this method.

However, it is not clear in this instance whether the differences in the results of the two methods are due to a failure of the two techniques to produce converging results (as discussed above, the allocation of point methodology may fail to take into account strength of preference or its explicit nature may be open to manipulation), or the fact that the techniques themselves elicit true preferences but they are different because the preferences are constructed in different ways. Other factors that may have influenced the results are design issues specific to this study rather than the techniques *per se*. The study focuses mainly on qualitative criteria which, even though given an ordered description, raises questions about the consistency of interpretation across individuals through the nature of the attribute levels. In the allocation of points exercise, respondents were presented with a list, as in shown in appendix 6; ranking the criteria in this way can influence (bias) respondents' ranking. Additionally, a list format causes problems for respondents having to divide points between so many criteria (six), and trying to think simultaneously of how many extra points one criterion deserves over another compared with all others may be unnecessarily complicated and alter the focus of the exercise. Respondents in this case were generally motivated to use simple numerical amounts that would emphasise a preference of one criterion over another but not necessarily indicate strength of preference or maintain it when compared with the other criteria. An exercise that involves the division of a pie chart into 'slices' representing criteria importance may overcome these problems.

In conclusion, it is recommended that further work should seek to develop these methodologies. Future work should also aim to incorporate the methodologies into the priority-setting process. In the UK (and, in fact, elsewhere) various frameworks are used to elicit and incorporate criteria weights and scores into priority-setting decision-making (see appendix 5). The examples, identified from the literature review, highlight three frameworks (or weighting and scoring methods) – rating exercises, allocation of points techniques and discrete choice experiments – that are commonly used for the purposes of scoring priorities. The first of these, assigning weights on a scale from 1 to *n*, has a number of limitations. First, there is no recognition of the need to make trade-offs between the criteria. Given this, it is possible that such a weighting exercise will result in high weightings being assigned to all criteria. Secondly, it is not clear whether such an approach indicates strength of preference (i.e. if one dimension is assigned a weight of 1 and another 5, does this mean that the latter is five times as important as the former)? Also, it has been shown that asking individuals to value dimensions of benefit individually will lead to different results to those arising from establishing weightings for the same dimensions when they are defined as a package.[475,531,588,589,590]

Ideally, a technique for weighting criteria should indicate not just the order of the weighting, but also the strength of preference for each criterion. Given that no one method is currently used in isolation or to its full potential, it is proposed that these limitations may be overcome by the adoption of either a discrete choice experiment or the allocation of points method[§], and future work should explore this.

Further, this study investigated the importance of community views in priority setting from the perspective of the policy-maker. However, an equally important perspective is that of the consumer and whether they want to be involved in priority-setting decisions (see chapter 8). The approach adopted in this chapter could easily be adapted to elicit public views concerning their involvement in decision-making. Experience regarding the eliciting of CHC views (another pilot study using this questionnaire indicated that CHC members had difficulty interpreting the meaning of the criteria) would suggest that alterations would be required to make it more understandable. Although the public does have the ability to answer choice-based CA questions if framed correctly (see chapter 5 for references applying to this technique), the public perspective should be a future area of research.

[§] See Farrar and colleagues[211] for an example of applying discrete choice experiments within a priority-setting framework.

Chapter 8

Discussion and conclusions

Results from the systematic review

Quantitative techniques

Quantitative techniques, classified as ranking, rating or choice-based approaches, were evaluated according to eight criteria: validity; reproducibility; internal consistency; acceptability to respondents; cost (financial and administrative); theoretical basis; whether the technique offered a constrained choice; and whether the technique provided a strength of preference measure.

Ranking exercises included simple ranking questions, QDP and CA. Simple ranking exercises have proved popular, probably because of the ease of both devising a ranking exercise and analysing the resulting data. However, the results of such exercises are of limited use. The QDP has not been used to date in healthcare. Given its ability to deal with the vagueness that exists in human decision-making, it may be particularly attractive in healthcare, where preferences are known to be both vague and difficult to articulate. CA ranking exercises provide useful information to policy-makers, and did well against the above criteria.

A number of rating scales were identified. The VAS has proved popular within the QALY paradigm. The main limitations of this technique are the lack of any constrained choice and the doubts expressed over whether the technique measures strength of preference. CA rating scales did well against the above criteria. A number of techniques were identified for eliciting attitudes, including Likert scales, the SDT and the Guttman scale. Satisfaction surveys have been frequently used to elicit public opinion. Their popularity probably reflects the relative ease of carrying out such surveys. Researchers should ensure that they construct sensitive techniques, or else use generic techniques where validity has already been established. Even when well-designed, sensitive techniques are used, users should be aware of the limited use of such techniques at the policy level. SERVQUAL appears to be a potentially useful technique and future research should consider its application in healthcare.

A number of choice-based techniques were identified. Three such techniques – MoV, AHP and allocation of points technique – have had limited application in healthcare, resulting in a small literature on their methodological status. Those more widely used in healthcare – SG, TTO, discrete choice CA and WTP – did well against the above criteria. Little methodological work is currently available on the PTO.

Qualitative techniques

Whilst chapter 5 defined a set of criteria for evaluating qualitative techniques, very few of the techniques identified for the review have been assessed according to these criteria. Although methodological issues can be extracted from the studies they are not explicitly addressed, and extracting them systematically has been challenging. Nonetheless, we believe that it was a useful process. We also conclude that it is necessary to continue the attempt to draw analogies between evaluations of quantitative and qualitative methods because users of the techniques should be able to choose either approach as best fits their purpose and context. It is also necessary to assess the several checklists set out in chapter 5 in respect of their ability to capture the criteria that we have used here.

In this review we attempted to assess qualitative techniques according to six criteria: validity; reliability; generalisability; objectivity; acceptability to respondents; and cost. All the methods were found to have distinct strengths and weaknesses, but there is a lot of ambiguity in the literature. There is disagreement about whether it is advantageous to use individual or group methods, but this very much depends on the specific topic being discussed and the people being asked. The most popular and widely used methods were one-to-one interviews and focus groups. In both of these methods it is crucial that the interviewer/moderator remains as objective as possible. It is impossible for the researcher to be completely objective but steps must be taken to minimise the opportunity for researcher bias. Both of these methods have potential problems with validity and reliability. It is the responsibility of the researcher to minimise these problems at all stages of data collection, analysis and

presentation of results. There were also many applications using the Delphi technique. In most of the examples listed experts or professionals participated but in some, however, patients were asked to take part; it is proposed that the Delphi technique could be more widely used to gain patients opinion. Citizens' juries were found to be very useful, especially when the subject matter is very complex. It is argued in the literature that as the participants have the opportunity for deliberation, their final decisions and opinions are more valid and reliable than an uninformed reaction to a question. There are problems with generalisability, however, because only very small numbers of people can be involved and it is very time-consuming and therefore costly. Consensus panels were found to have similar methodological considerations to citizens' juries but have been criticised for not allowing sufficient time for participants to make decisions. They are, however, less costly and cannot be dismissed at this stage. Public meetings are frequently used and are a quick and inexpensive way of gaining public opinion. It has been argued, however, that they are unrepresentative and therefore have limited value. Complaints procedures are noted but will only consider particularly negative view points, and have limited value in assessing public opinion. Three additional methods are included: nominal group technique; concept mapping; and case study analysis. It is proposed that they may be useful, but further research is needed to test their use for issues in priority setting.

Results from primary research

Both the choice-based CA technique and the allocation of points method found the views of the public to be important in the priority-setting exercise, although the relative rankings differed across the two techniques. In the follow-up telephone interviews, whilst the majority of respondents stated that the community had a role to play in decision-making, and that this role was important or very important in the context of priority setting, they ranked this criterion as the least important of the six.

General issues raised

Although not the main concern of this study, the literature searches highlighted a number of more general issues that will be raised as attempts are made to elicit public preferences for use in the priority-setting exercise. These are independent of the techniques employed and include the following.

Who are the public?

An important question within the context of eliciting public preferences for use in priority setting is whose values should be used? Should we obtain values from actual users of the service or should we elicit the views of the community more generally? There is no agreed answer to this question. Gafni[591] argued that, within the context of a publicly provided healthcare system, it is the views of the community that are relevant. Shackley and Ryan,[396] however, argue that the answer to this question depends on the context of the question being addressed. If policy is concerned with what health services should be provided (hearts versus hips versus helicopters) then it is the views of the community that are relevant. However, if questions are concerned with how to provide, then the preferences of patients are relevant.

Do the public want to be involved in healthcare decision-making?

An implicit assumption so far has been that the public want their views to be considered. In a recent review of published studies, Benbassat and colleagues[592] concluded that whilst most patients want to be informed about their illness, only a proportion want to be involved in planning their care, and that some patients would prefer to be completely passive. Determinants of a desire to be passive include medical condition and socio-economics factors. Similar conclusions were reached by Guadagnoli and Ward.[593]

In a survey conducted by the British Medical Association and The King's Fund,[594] only 22% of the general public thought that they should make prioritisation decisions. However, in this study the level of decision-making was restricted to prioritising amongst individuals. In contrast, in a large study conducted by Bowling,[595] based on a random sample of the British population, it was found that most respondents (88%) thought that surveys of public opinion should be used in planning services. Groves,[17] in the context of the British Medical Association survey, found that one-third of the public sample thought the public should have a say in the process. This compared to half the managers and one-fifth of the doctors. In the qualitative literature, Coote and Lenaghan[331] reported on a series of pilot citizens' juries. One of the juries addressed the question of whether the public should be involved in priority setting. This jury concluded that the public should be involved in decision-making in conjunction with other experts. Across all five citizens' juries, jurors noted that the importance of public involvement was the main motivational

reason for their involvement. The authors concluded that jurors were willing to be involved in the decision-making process, although financial remuneration may be necessary to compensate for their time, and it should be remembered that juries are to compose only one part of the decision-making process. Likewise, Lenaghan and colleagues[332] observed that given "enough time and information" the public is "willing and able" to contribute to the priority-setting process. Kneeshaw,[596] reviewing public involvement in healthcare, reported that the public wanted to be involved in the decision-making process, although the final decisions should be left to doctors. Related to this, there is also some evidence that there is disutility associated with patient involvement in the decision-making process.[597,598]

Do public preferences for healthcare exist?

Researchers concerned with eliciting public preferences in the delivery of healthcare make an implicit assumption that such preferences exist. Within psychology, Fischoff[599] made a distinction between the **philosophy of articulated values** and the **philosophy of basic values**. The former represents a state where individuals have well-articulated values, which the researcher can extract. The latter holds that individuals lack well-formed preferences for all but the most familiar of goods or services. This familiarity is a result of having made numerous choices in the past (and errors), such that a complete preference structure can be settled on. In reality, individuals' preferences for any given good or service may lie somewhere on a continuum between these two philosophies. Psychologists have also developed the notion of 'constructive nature of preferences' to explain preference reversals in experimental studies.[478,600,601] Here it is argued that individuals 'construct' their preferences as they answer experimental questions.

Given the nature of the commodity healthcare, preferences for this good may not be complete. Individuals often do not have experience of healthcare interventions. There is an asymmetry of information and agents often make decisions on behalf of patients and the public. There is also, in many publicly provided healthcare systems, a lack of choice*. Thus, healthcare may be better placed under a 'basic values' philosophy. Shiell and colleagues[602] noted the possibility of people not having well-formed preferences for health and healthcare, arguing that the values elicited from experimental studies may not equate with actual preferences. However, no empirical work was carried out. Dolan and colleagues[323] found that that respondents changed their views concerning priority setting in healthcare between two questionnaires, and argued that this change in preferences was due to discussions held between the questionnaires. (See page 68.)

Such findings have implications for the way research is carried out into public preferences concerning healthcare. If respondents really do construct their preferences as they go along, then 'valid' results may only be obtained after successive surveys (quantitative or qualitative) or interactive discussions have taken place.

Clarke[567] advocates the use of deliberation in finding out public opinion. It is believed that if people have the chance to deliberate then their judgements and conclusions will be more valid than mere gut reactions to a question. The deliberation process involves participants reasoning, reflecting and refining their ideas, all of which can lead to an informed and committed decision. There are various degrees of deliberation in the methods previously described in this review. Clarke states that focus group discussion involves a low level and citizens' juries involve a high level, and that with increased levels of deliberation comes increased cost.

Clarke[567] proposes a number of methods to help the deliberation process. First, community issue groups. Such groups lie between focus groups and citizens' juries on the scale of deliberation. They consist of groups of 8–12 people who meet several times over a set period of time. These groups then join other similar groups to form a network that is committed to discussing common issues. Secondly, community workshops, which have a variable level of deliberation, are more in-depth than a focus group but can obtain a larger sample than a citizens' jury. These consist of 12–20 people considered to be representative of the wider community. They meet for 1 day to discuss a particular issue with the aid of a moderator. Recommendations are made to a commissioning body. This method is subject

* Whilst people in a publicly provided healthcare system do not have much experience with choosing between healthcare interventions and paying for them, such an argument may not apply in privately provided healthcare systems. Thus, in such systems, preferences may be more complete and stable. This is ultimately an empirical issue.

to the same potential problems as some of the other techniques; that is, problems of gaining a representative sample and ensuring the use of a skilled and unbiased moderator.

The Sedgefield Health Alliance formed a local advisory group to help the deliberation process.[603] It was set up between the County Durham Health Authority, the Sedgefield Borough Council and Durham County Social Services. The aim of joining these groups together was to improve the health and quality of life of people living and working in the area. It was proposed that this be used as a method of enabling the exchange of information and debate of relevant issues and to raise the profile of health and social issues and to stimulate local interest in these issues. A total of 19 meetings were held during the period April to June 1999. A variety of deliberative methods were used during these meetings and a number of concerns and issues were identified by the participants. As a means of validation, the public requested a report summarising the topics raised and asked that information should be fed back to them at organised liaison meetings. The authors state that this method's success lies in the conviction of board members to appreciate and respond to the issues raised in the meetings – an assessment of the outcome rather than of the process of consultation.

Quantitative and qualitative techniques may be used together to help the deliberation process. Dolan and colleagues[323] conducted a series of structured focus groups to investigate peoples' views about priority setting and how these views change after discussions in the groups. The authors randomly selected 72 people from six age–sex categories. Participants were split into ten groups and told they would be paid £30 for attending after the conclusion of the second meeting. The discussion groups were moderated by two of the researchers who defined the agenda and gave each group member the opportunity to speak. Two meetings with each group were conducted, separated by 2 weeks. The participants were asked to fill out a questionnaire at the beginning of the first meeting and at the end of the second to determine how their views had altered after the opportunity for deliberation. The questionnaire consisted of two groups of questions. First, participants were asked whether certain groups (e.g. doctors and nurses, politicians) should have more or less involvement in making decisions about prioritising healthcare. Secondly, they were asked whether a specified group of people (e.g. children, disabled, married people) should have a higher or lower priority for treatment. Both of these questions used a four-point Likert scale. Participants were also, at the start of the second meeting, asked to prioritise between four people – all of whom needed treatment for different reasons. The authors found that participants' views changed after the two meetings to the extent that they no longer discriminated quite so heavily against smokers, drinkers and drug takers. This study shows that quantitative research can be cultivated from focus groups.

Preferences for the status quo

With regard to eliciting public preferences regarding new health technologies, there is some evidence that people will adopt a conservative response, preferring the *status quo*.[61,439,604–606] Cartwright[604] found that within the context of maternity care, women tend to choose the options they have experienced. Similarly, Porter and MacIntyre,[439] in a study concerned with innovations in antenatal care, found that pregnant women preferred the type of care they received. Ryan and colleagues[61] found that, within the context of a RCT, individuals who experienced the introduction of a patient health card valued it, whereas those who had no experience did not value it. Bate and Ryan[605] argued that preferences for junior doctor-led care over nurse-led care in the provision of rheumatology services could be partly explained by a lack of experience with the latter.

This decision-making heuristic has important implications in a climate where public views are being promoted within the context of priority setting. The logical implication is that if the public prefer what they know, new innovations may not be valued sufficiently to persuade them to change from the *status quo*.

Explanations for preference for the *status quo* has received attention in the economics literature. This anomaly in choice is known as the **endowment effect**[607] (also termed *status quo* bias[608]) and refers to a situation whereby people value goods more highly once they own them (or have experience of them). This results in individuals often demanding much more to give up an object than they would be willing to pay to receive it. The endowment effect has been offered as an explanation for the well-documented discrepancies between WTP and WTA (willingness to accept) in the contingent valuation literature.[609–613]

Another hypothesis for such conservative preferences is that individuals are attempting

to minimise the potential to experience the psychological feelings of regret and disappointment. In explaining consistent violations of EUT, Bell[476,480] and Loomes and Sugden[477,481] argue that when making decisions, individuals take account not only of the final outcome but also of the chance of experiencing regret and disappointment. As indicated above, evidence has suggested that regret is an important element in individual valuation and decision-making in healthcare.[482,483] The potential to experience both regret and disappointment will be minimised if individuals choose the *status quo*.

Kahneman and Tversky argue that that **loss aversion** may help to explain violations of EUT.[475] Here, a reference point, usually the *status quo*, determines that individual preferences and the disutility of giving up an object is greater than the utility associated with acquiring it. The anomalies of regret and the endowment effect both manifest themselves within the concept of **loss aversion**.

Within the health service research literature, a limited amount of work has attempted to explain preferences for the *status quo*. In attempting to explain why pregnant women prefer the care they receive, Porter and MacIntyre[439] argued that patients assume that whatever service is offered has been carefully considered by experts and is therefore likely to be the best for them. That is, 'what is, must be best'. Such beliefs naturally result in aversion to innovations. Ryan and colleagues[61] and Bate and Ryan[605] both argued that the tendency to favour the *status quo* may be explained by a lack of information about the alternative. Under such circumstances, the potential for *status quo* bias will be increased if individuals are risk averse.

Ethical issues in involving the public

Assuming that the public (however defined) want to be involved, that they have well-defined preferences, and that techniques are used to overcome potential problems of preferences for the *status quo* (as well as satisfying criteria for a good technique), is public involvement a good thing? Involving the public in the planning and prioritising of services is fraught with ethical as well as practical issues. One major issue that will be encountered is how to deal with conflicting views about priorities (both between subgroups of the population as well as the public and health professionals). Potential differences in opinion imply that important ethical issues will be raised as attempts are made to incorporate public opinion into priority setting. For example, Stronks and colleagues,[336] in a study using consensus panels, observed that patients and health insurers were acting in their own interests. The same could not be said of the public or healthcare professionals (see also Dicker and Armstrong[289]). The public favoured individual responsibility (i.e. co-payments) without any consideration of the groups this may exclude. This strategy may threaten the concept of equal access for equal need. In contrast, professionals took into account the consequences of their decisions on equality of access. Stronks and colleagues[336] concluded that

> "it is not clear that including all the different actors in the decision-making process of prioritisation of health services will lead to more equitable or broadly supported outcomes or to better health for the population ... their decisions might threaten the universal accessibility of core services".

Studies identified in our review found that the public have set views over who they would prefer to see treated, with females being preferred to males, non-smokers to smokers, poor people to rich, non-drinkers over drinkers, whether British born, Christians over atheists.[18,25,26,68] Issues will inevitably be raised here concerning the extent to which policy-makers take such views into account when allocating scarce resources. What if policy-makers do not like the views of the public?

Overcoming tokenism

It is important that the public believe that their views are going to be incorporated into the priority-setting process, and do not view such elicitation exercises as tokenism. 'Planning for real' has been proposed as a method to ensure this.[614] This method was originally devised to aid people in planning their physical environment and surroundings in a user-friendly fashion. Klein has suggested that the theory of this method transfers very well in a health context.[614] It involves consulting the public at all levels related to issues such as who to involve, what level of involvement to seek, who to consult, what the task is and what type of support is needed to achieve this task. This will ensure that the public take the exercise seriously. Also advocated is the encouragement of people to play an active role in creating new ideas, plans and projects. However, 'Planning for real' has not been evaluated.

Bringing together quantitative and qualitative techniques

Many applied health (service) researchers argue that qualitative and quantitative methods can and should complement each other. Pope and Mays[157]

highlighted three ways in which this could be done:

- First, many quantitative researchers are familiar with the idea of using qualitative methods in order to prepare their quantitative techniques; for example, using interviews and/or focus groups to determine the questions for a postal questionnaire, the content of an outcome indicator, or the criteria to include in a CA study.
- Secondly, qualitative techniques can be used in parallel with quantitative ones, either (1) to help explain quantitative findings; (2) to enlarge on such findings; or (3) as part of triangulation.[615] Whilst quantitative techniques tell us that the public prefer local clinics or do not want resources to be allocated to certain social groups, qualitative methods can help us understand the reasons behind such preference structures. Using both types of techniques in parallel can be part of a validation process, as in a triangulation exercise whereby the different techniques address the same issue from a slightly different direction.
- Finally, qualitative methods may be used to explore "complex phenomena or areas not amenable to quantitative research".[157] An example of this would be using qualitative techniques to establish the cognitive strategies and decision-making heuristics employed when responding to quantitative questionnaires.

Chapter 9

Recommendations on the use of techniques and future research

Recent UK Government policy suggests an increased role for the public in priority setting. The purpose of this review has been to identify and evaluate techniques for eliciting public preferences, and to assess the importance of public preferences *vis-à-vis* other criteria for setting priorities. Consideration was also given to general issues that were identified in the literature that will be raised as public preferences are elicited for use in priority setting. This final chapter includes recommendations on what techniques to use. Finally, recommendations for future research are proposed.

Guidance on the use of techniques

It is concluded that there is no single best method to gain public opinion. There are no hard and fast rules on which method should be used to answer a certain question, but the method must be carefully chosen and rigorously carried out in order to accommodate the question being asked. Authors should be encouraged to set out their critiques more systematically.

Quantitative techniques

Based on the evidence currently available, for the ranking approaches, we recommend that researchers consider the QDP and CA. From the rating approaches, CA should be further explored. Satisfaction surveys may be a useful technique to measure outcomes and to assess consultations and patterns of communication. Researchers using such surveys should ensure that they construct sensitive techniques, or else use generic techniques where validity has already been established. Even when well-designed, sensitive techniques are used, users should be aware of the limited use of such techniques. SERVQUAL appears to be a potentially useful technique and future research should consider its application in healthcare. Likert scales, Guttman scales and the SDT all provide useful information on attitudes. All the choice-based techniques should be (further) explored.

Qualitative methods

Several qualitative methods were identified which should be considered when eliciting public preferences. The technique used will depend on the question being asked and also the types of people being asked. Individual interviews and focus groups are particularly useful in gaining detailed data from a limited number of individuals; these may be the best ways of looking into sensitive topics. The Delphi process, thus far mainly used in eliciting responses from professionals, could be more extensively used to gain the public/patient perspective. Citizens' juries have been shown to be extremely useful in looking at complex issues but are very costly and require a lot of preparation and planning. Public meetings are a quick and cheap method of gaining a feel for public opinion but may only take very extreme views into consideration. Further work may need to be conducted to test the value of using consensus panels in this context, and complaints procedures have limited value because they only consider very negative views of healthcare. Additionally, further work may be needed to establish the appropriateness of using the special analysis techniques of nominal group technique, concept mapping and case study analysis.

Future research

The research agenda is broken down into research questions relating to the systematic review of techniques and those relating to more general issues that were raised whilst conducting the review. The research agenda related to the cognitive strategies and decision-making heuristics that respondents adopt when completing quantitative surveys should be seen as a priority. The general issues emerging should be given equal priority.

Researching techniques:

- the techniques recommended above should continue to be researched
- research to investigate the AHP, MoV, allocation of points, QDP, SERVQUAL and PTO as quantitative methods, and telephone, email and dyadic interview techniques, consensus panels, case study analyses, concept mapping and nominal group techniques as qualitative methods

- when addressing the above points, quantitative and qualitative methods should be used more alongside each other. A crucial part of this research agenda should address the extent to which preferences for healthcare exist, as well as the cognitive strategies and decision-making heuristics respondents adopt when completing quantitative surveys. This should involve extensive piloting work using qualitative approaches to inform the design and interpretation of quantitative studies.

General issues raised in the review:

- do the public want to be involved in healthcare decision-making?
- potential problems encountered with a preference for the *status quo*
- ethical issues in involving the public
- developments of frameworks to ensure public preferences are incorporated into priority setting.

The recommendation to apply qualitative research to look at the cognitive strategies and decision-making heuristics individuals employ when answering such questionnaires will have implications for whether the techniques should be used to elicit public preferences. Therefore, we would recommend that this review of quantitative instruments for eliciting preferences be updated to take account of work done in this field over the coming years. It is difficult to predict at the moment how quickly this research will be conducted. Given the nature of such research, it is unlikely that a substantial body of literature will exist in the next couple of years, and such an update should be considered in 5–10 years time.

Acknowledgements

This study was commissioned by the NHS R&D Executive's Health Technology Assessment programme, project number 96/49/04.

This work is the product of a collaboration among eight authors. Moira Napper was responsible for the identification of the techniques within the systematic review. She was helped by David Scott for the quantitative methods and Caroline Reeves for the qualitative methods. Mandy Ryan and David Scott carried out the review of quantitative techniques and Caroline Reeves the review of qualitative techniques. Edwin van Teijlingen helped with the review of the SDT, Guttman scales and satisfaction surveys and Christian Robb with the review of techniques for visual analogue, SG and TTO. Angela Bate and Mandy Ryan carried out the primary research concerned with the importance of public views in decision-making and these authors, together with David Scott, carried out the primary research looking at the weights of the methodological criteria. Mandy Ryan took overall responsibility for writing the report. Edwin van Teijlingen and Elizabeth Russell made invaluable comments to all aspects of the report, and were involved in the writing up. The views expressed in this report are those of the authors, who are also responsible for any errors.

The authors would like to thank Penelope Mullen, Shirley McIver (both from the University of Birmingham) and Shelley Farrar (Health Economics Research Unit, University of Aberdeen) for their comments on earlier drafts. Thanks also go to Anne Bews, Flora Buthlay, Margaret Beveridge and Sophie Davidson for their help with the photocopying and mailing of the questionnaires. The authors are also indebted to all the respondents to the questionnaire-based surveys and to Andrew Walker (Greater Glasgow Health Board) for his help with the primary research assessing the importance of public preferences in priority setting. Finally, thanks to the anonymous referees who provided invaluable comments and HTA series editors for their editing, proof-reading and, above all, perseverance.

References

1. NHS Management Executive. Local voices: the views of local people in purchasing for health. London: NHS Management Executive; 1992.
2. Working for patients. Command Paper 555. London: HMSO; 1989.
3. Promoting better health. Command Paper 249. London: HMSO; 1989.
4. The health of the nation. Command Paper 1986. London: HMSO; 1991.
5. NHS Centre for Reviews & Dissemination. Undertaking systematic reviews of research on effectiveness: CRD guidelines for those carrying out or commissioning reviews. York: NHS CRD, University of York; 1996. NHSCRD Report no. 4.
6. Black N, Brazier J, Fitzpatrick R, Reeves B, editors. Health services research methods: a guide to best practice. London: BMJ Books; 1998.
7. Bryman A. Quantity and quality in social research. London: Routledge; 1993.
8. Baker TL. Doing social research. 2nd ed. New York: McGraw-Hill; 1994.
9. Bowling A. Research methods in health: investigating health and health services. Buckingham: Open University Press; 1997.
10. Flick U. An introduction to qualitative research. London: Sage; 1998.
11. Denzin NK. The research act: a theoretical introduction to sociological methods. New York: McGraw-Hill; 1978.
12. Sapsford R, Abbott P. Research methods for nurses and the caring professions. Buckingham: Open University Press; 1992.
13. Mays N, Pope C. Qualitative research in health care: assessing quality in qualitative research. *BMJ* 2000;**320**:50–2.
14. Edwards SJL, Lilford RJ, Kiauka S. Different types of systematic review in health services research. In: Black N, Brazier J, Fitzpatrick R, Reeves B, editors. Health services research methods: a guide to best practice. London: BMJ Books; 1998. p. 255–9.
15. Hutton JL, Ashcroft R. What does "systematic" mean for reviews of methods? In: Black N, Brazier J, Fitzpatrick R, Reeves B, editors. Health services research methods: a guide to best practice. London: BMJ Books; 1998. p. 249–54.
16. Bowling A, Jacobson B, Southgate L. Explorations in consultation of the public and health professionals on priority setting in an inner London health district. *Soc Sci Med* 1993;**37**:851–7.
17. Groves T. Public disagrees with professionals over NHS rationing [News]. *BMJ* 1993;**306**:673.
18. Furnham A, Meader N, McClelland A. Factors affecting nonmedical participants' allocation of scarce medical resources. *J Soc Behav Pers* 1998;**13**:735–46.
19. Widmark-Petersson V, von Essen L, Sjoden PO. Cancer patient and staff perceptions of caring and clinical care in free versus forced choice response formats. *Scand J Caring Sci* 1998;**12**:238–45.
20. Jacobson B, Bowling A. Involving the public: practical and ethical issues [review]. *Br Med Bull* 1995;**51**:869–75.
21. Rosko MD, McKenna W. Modeling consumer choices of health plans: a comparison of two techniques. *Soc Sci Med* 1983;**17**:421–9.
22. Kinnunen J, Lammintakanen J, Myllykangas M, Ryynanen OP, Takala J. Health care priorities as a problem of local resource allocation. *Int J Health Plann Manage* 1998;**13**:216–29.
23. Ryynanen OP, Myllykangas M, Niemela P, Kinnunen J, Takala J. Attitudes to prioritization in selected health care activities. *Scand J Soc Welfare* 1998;**7**:320–9.
24. Angermeyer MC, Matschinger H, RiedelHeller SG. Whom to ask for help in case of a mental disorder? Preferences of the lay public. *Soc Psychiatry Psychiatr Epidemiol* 1999;**34**:202–10.
25. Furnham A. Factors relating to the allocation of medical resources. *J Soc Behav Pers* 1996;**11**:615–24.
26. Furnham A, Briggs J. Ethical ideology and moral choice: a study concerning the allocation of medical resources. *J Soc Behav Pers* 1993;**8**:87–98.
27. Tolley K, Whynes D. Rationing in the NHS: setting priorities depends on who you ask and how you ask it. Paper presented to the UK Health Economists' Study Group meeting; 1994 Jul; Newcastle.
28. Bryson N, Ngwenyama OK, Mobolurin A. A qualitative discriminant process for scoring and ranking in group support systems. *Information Process Manage* 1994;**30**:389–405.
29. Bryson N. Supporting consensus formation in group support systems using the qualitative discriminant process. *Ann Oper Res* 1997;**71**:75–91.

30. Ngwenyama OK, Bryson N. Generating belief functions from qualitative preferences: an approach to eliciting expert judgments and deriving probability functions. *Data Knowledge Eng* 1998;**28**:145–59.

31. Lancaster K. Consumer demand: a new approach. New York: Columbia University Press; 1971.

32. Luce D, Tukey J. Simultaneous conjoint measurement: a new type of fundamental measurement. *J Math Psycho* 1964;**1**:1–27.

33. Cattin P, Wittink D. Commercial use of conjoint analysis: a survey. *J Marketing* 1982;**46**:44–53.

34. *Journal of Transport Economics and Policy* 1988;**22**(1).

35. Rae D. Visibility impairment at Mesa Verde National Park: an analysis of benefits and costs of controlling emissions in the Four Corners area. Report prepared for Electric Power Research Institute. Boston, MA: Charles River Associates; 1981.

36. Rae D. Benefits of improving visibility at Great Smoky National Park. Draft report prepared for Electric Power Research Institute. Boston, MA. Charles River Associates; 1981.

37. Chinburapa V, Larson LN, Brucks M, Draugalis J, Bootman JL, Puto CP. Physician prescribing decisions: the effects of situational involvement and task complexity on information acquisition and decision making. *Soc Sci Med* 1993;**36**:1473–82.

38. Gelb GM, Gelb BD. Physicians and hospital decision making: a two-stage technique for improvement. *Hosp Health Serv Adm* 1987;**32**:139–49.

39. McClain JO, Rao VR. Trade-offs and conflicts in evaluation of health system alternatives: methodology for analysis. *Health Serv Res* 1974;**9**:35–52.

40. Nickerson CA, McClelland GH, Petersen DM. Measuring contraceptive values: an alternative approach. *J Behav Med* 1991;**14**:241–66.

41. Orkin FK, Greenhow DE. A study of decision making. *Anesthesiology* 1978;**48**:267–71.

42. Parker BR, Srinivasan V. A consumer preference approach to the planning of rural primary health-care facilities. *Oper Res* 1976;**24**:991–1025.

43. Rosko MD, Walker LR, McKenna W, DeVita M. Measuring consumer preferences for ambulatory medical care arrangements. *J Med Syst* 1983;**7**:545–54.

44. Shemwell DJ, Yavas U. Congregate care facility selection: a conjoint approach. *Health Marketing Q* 1997;**14**(4):109–20.

45. Carroll NV, Gagnon JP. Consumer demand for patient-oriented pharmacy services. *Am J Public Health* 1984;**74**:609–11.

46. Singh J, Cuttler L, Shin M, Silvers JB, Neuhauser D. Medical decision-making and the patient: understanding preference patterns for growth hormone therapy using conjoint analysis. *Med Care* 1998;**36** Suppl 8:AS31–45.

47. Stevens SS. A metric for social consensus. *Science* 1966;**151**:530–41.

48. Armstrong RA, Brickley MR, Shepherd JP, Kay EJ. Healthy decision making: a new approach in health promotion using health state utilities. *Community Dent Health* 1995;**12**:8–11.

49. Brazier J, Deverill M. The use of health related quality of life measures in economic evaluation. In: Black N, Brazier J, Fitzpatrick R, Reeves B, editors. Health services research methods: a guide to best practice. London: BMJ Books; 1998. p. 23–34.

50. Williams A. Economics of coronary bypass grafting. *BMJ* 1985;**291**:326–9.

51. Bunch WH, Chapman RG. Patient preferences in surgery for scoliosis. *J Bone Joint Surg Am* 1985;**67**:794–9.

52. Chakraborty G, Gaeth GJ, Cunningham M. Understanding consumers' preferences for dental service [review]. *J Health Care Mark* 1993;**13**(3):48–58.

53. Chinburapa V, Larson LN. Predicting prescribing intention and assessing drug attribute importance using conjoint analysis. *J Pharm Mark Manage* 1988;**3**(2):3–18.

54. Diamond JJ, Ruth DH, Markham FW, Rabinowitz HK, Rosenthal MP. Speciality selections of Jefferson Medical College students. *Educ Health Prof* 1994;**17**:322–8.

55. Geiger CJ, Wyse BW, Parent CR, Hansen RG. Nutrition labels in bar graph format deemed most useful for consumer purchase decisions using adaptive conjoint analysis. *J Am Diet Assoc* 1991;**91**:800–7.

56. Graf MA, Tanner DD, Swinyard WR. Optimizing the delivery of patient and physician satisfaction: a conjoint analysis approach. *Health Care Manage Rev* 1993;**18**(4):34–43.

57. Harrison DD, Cooke CW. An elucidation of factors influencing physicians' willingness to perform elective female sterilization. *Obstet Gynecol* 1988;**72**:565–70.

58. Harwood RH, Rogers A, Dickinson E, Ebrahim S. Measuring handicap: the London Handicap Scale, a new outcome measure for chronic disease. *Qual Health Care* 1994;**3**:11–16.

59. Reardon G, Pathak DS. Segmenting the antihistamine market: an investigation of consumer preferences. *J Health Care Mark* 1990;**10**(3):23–33.

60. Ryan M, Shackley P, McIntosh E. Using conjoint analysis to elicit the views of health service users: an application to the patient health card. *Health Expectations* 1998;**1**:117–29.

61. Ryan M, McIntosh E, Shackley P. Methodological issues in the application of conjoint analysis in health care. *Health Econ Lett* 1998;**7**:373–8.

62. Szeinbach SL, Mason HL, Schondelmeyer SN, Collins PD. Variables affecting pharmacists' willingness to accept third-party prescription contracts: a conjoint analysis. *J Health Care Mark* 1990;**10**(3):45–50.

63. van der Pol M, Ryan M. Using conjoint analysis to establish consumer preferences for fruit and vegetables. *Br Food J* 1996;**98**(8):5–12.

64. Wigton RS, Hoellerich VL, Patil KD. How physicians use clinical information in diagnosing pulmonary embolism: an application of conjoint analysis. *Med Decis Making* 1986;**6**:2–11.

65. Wind Y, Spitz LK. Analytical approach to marketing decisions in health-care organisations. *Oper Res* 1976;**24**:973–90.

66. Richardson DK, Gabbe SG, Wind Y. Decision analysis of high-risk patient referral. *Obstet Gynecol* 1984;**63**:496–501.

67. Coen R, O'Mahoney D, O'Boyle C, Joyce CRB, Hiltbrunner B, Walsh JB, *et al.* Measuring the quality of life of dementia patients using the schedule for the evaluation of individual quality of life. *Ir J Psychol* 1993;**14**:154–63.

68. O'Boyle CA, McGee H, Hickey A, O'Malley K, Joyce CRB. Individual quality of life in patients undergoing hip replacement. *Lancet* 1992;**339**:1088–91.

69. Browne JP, O'Boyle A, McGee HM, Joyce RB, McDonald NJ, O'Malley K, *et al.* Individual quality of life in the healthy elderly. *Qual Life Res* 1994;**3**:235–44.

70. McGee HM, O'Boyle CA, Hickey A, O'Malley K, Joyce CRB. Assessing the quality of life of the individual: the SEIQoL with a healthy and a gastroenterology unit population. *Psychol Med* 1998;**21**:749–59.

71. Oppenheim AN. Questionnaire design, interviewing and attitude measurement. London: Pinter; 1992.

72. Aldrich JH, Niemi RG, Rabinowitz G, Rohde DW. The measurement of public-opinion about public-policy – a report on some new issue question formats. *Am J Political Sci* 1982;**26**:391–414.

73. Richardson A, Charny M, Hanmerlloyd S. Public-opinion and purchasing. *BMJ* 1992;**304**:680–2.

74. Lahti S, Tuutti H, Hausen H, Kaariainen R. Opinions of different subgroups of dentists and patients about the ideal dentist and the ideal patient. *Community Dent Oral Epidemiol* 1995;**23**:89–94.

75. Weekman VW. Laboratory reactors and their limitations. *Am Instit Chem Eng J* 1974;**20**:833–40.

76. Borowsky SJ, Davis MK, Goertz C, Lurie N. Are all health plans created equal? The physician's view. *JAMA* 1997;**278**:917–21.

77. Tymstra T, Andela M. Opinion of Dutch physicians, nurses and citizens on health care policy, rationing and technology. *JAMA* 1993;**270**:2995–9.

78. Lahti S, Tuutti H, Hausen H, Kaariainen R. Comparison of ideal and actual behavior of patients and dentists during dental treatment. *Community Dent Oral Epidemiol* 1995;**23**:374–8.

79. Lahti S, Tuutti H, Hausen H, Kaariainen R. Dentist and patient opinions about the ideal dentist and patient – developing a compact questionnaire. *Community Dent Oral Epidemiol* 1992;**20**:229–34.

80. Marks GB, Dunn SM, Woolcock AJ. A scale for the measurement of quality-of-life in adults with asthma. *J Clin Epidemiol* 1992;**45**:461–72.

81. Conway T, Hu TC, Harrington T. Setting health priorities: community boards accurately reflect the preferences of the community's residents. *J Community Health* 1997;**22**:57–68.

82. Checkoway B. Public participation in health planning agencies: promise and practice. *J Health Polit Policy Law* 1982;**7**:723–33.

83. Pfennings L, Cohen L, Vanderploeg H. Preconditions for sensitivity in measuring change – visual analog scales compared to rating-scales in a likert format. *Psychol Rep* 1995;**77**:475–80.

84. Wilson JR, Fazey JA. Self-esteem, compliance, and cervical screening. *Psychol Rep* 1995;**77**:891–8.

85. Holmes C. A statistical evaluation of rating scales. *J Mark Res Soc* 1974;**16**:87–107.

86. McKennell AC, Bryner JN. Self images and smoking behaviour among school boys. *Br J Educ Psychol* 1969;**39**:27–39.

87. Valois P, Godin G. The importance of selecting appropriate adjective pairs for measuring attitude based on the semantic differential method. *Qual Quantity* 1991;**25**:57–68.

88. Wilbur J, Miller A, Montgomery A. The influence of demographic characteristics, menopausal status, and symptoms on women's attitudes toward menopause. *Women Health* 1995;**23**:19–39.

89. Bowles C. Measure of attitude toward menopause using the semantic differential model. *Nursing Res* 1986;**35**:81–5.

90. Wikblad KF, Wibell LB, Montin KR. The patient's experience of diabetes and its treatment – construction of an attitude scale by a semantic differential technique. *J Adv Nurs* 1990;**15**:1083–91.

91. Nichols BS, Misra R, Alexy B. Cancer detection: How effective is public education? *Cancer Nurs* 1996;**19**:98–103.

92. Swain R, McNamara M. The effects of a participative programme on Irish pupils' attitudes to HIV/AIDS. *Health Educ Res* 1997;**12**:267–73.

93. Girón M, Gomezbeneyto M. Relationship between family attitudes measured by the semantic differential and relapse in schizophrenia – a 2-year follow-up prospective-study. *Psychol Med* 1995;**25**:365–71.

94. Wierenga B, Ophuis PAMO, Huizingh EKR, Vancampen PAFM. Hierarchical scaling of marketing decision-support systems. *Decis Support Syst* 1994;**12**:219–32.

95. Kelloway EK, Barling J. Members' participation in local union activities – measurement, prediction, and replication. *J Appl Psychol* 1993;**78**:262–79.

96. Katz YJ, Schmida M. A guttman scale factor structure of comprehensiveness. *Educ Psychol Meas* 1993;**53**:225–32.

97. Podell L, Perkins JC. A Guttman scale for sexual experience – a methodological note. *J Abnorm Psychol* 1957;**54**:420–2.

98. Edmundson EW, Koch WR, Silverman S. A facet analysis approach to content and construct-validity. *Educ Psychol Meas* 1993;**53**:351–68.

99. Santos ML, Booth DA. Influences on meat avoidance among British students. *Appetite* 1996;**27**:197–205.

100. Sanner M. A comparison of public-attitudes toward autopsy, organ donation, and anatomic dissection – a Swedish survey. *JAMA* 1994;**271**:284–8.

101. Oshea E, Murray P. Care provision and dependency in long-stay institutions. *Econ Soc Rev* 1997;**28**:43–61.

102. Petersen PE. Guttman scale analysis of dental-health attitudes and knowledge. *Community Dent Oral Epidemiol* 1989;**17**:170–2.

103. Khayat K, Salter B. Patient satisfaction surveys as a market-research tool for general practices. *Br J Gen Pract* 1994;**44**:215–19.

104. Ware JE, Hays RD. Methods for measuring patient satisfaction with specific medical encounters. *Med Care* 1988;**26**:393–402.

105. Makhdoomn YM, Elzubier AG, Hanif M. Satisfaction with health care among primary health care centers attendees' in Al-Khobar, Saudi Arabia. *Saudi Med J* 1997;**18**:227–30.

106. Chiu L. Family caregiver's satisfaction with home care in the Taipei Metropolitan Area. *Public Health Nurs* 1997;**14**:42–50.

107. Sharma RD, Chahal H. Patient satisfaction in public health system – a case study. *Indian J Soc Work* 1995;**56**:445–56.

108. Cohen G, Forbes J, Garraway M. Can different patient satisfaction survey methods yield consistent results? Comparison of three surveys. *BMJ* 1996;**313**:841–4.

109. Grogan S, Conner M, Willits D, Norman P. Development of a questionnaire to measure patients' satisfaction with general practitioners' services. *Br J Gen Pract* 1995;**45**:525–9.

110. Loeken K, Steine S, Sandvik L, Laerum E, Finset A. A new measure of patient satisfaction with mammography. Validation by factor analytic technique. *Fam Pract* 1996;**13**:67–74.

111. Avis M, Bond M, Arthur A. Questioning patient satisfaction: an empirical investigation in two outpatient clinics. *Soc Sci Med* 1997;**44**:85–92.

112. Counte MA. An examination of the convergent validity of three measures of patient satisfaction in an outpatient center. *J Chronic Dis* 1979;**32**:583–8.

113. Finkelstein BS, Harper DL, Rosenthal GE. Patient assessments of hospital maternity care: a useful tool for consumers? *Health Serv Res* 1999;**34**:623–40.

114. Ruggeri M, Dall'Agnola R. The development and use of the Verona Expectations for Care Scale (VECS) and the Verona Service Satisfaction Scale (VSSS) for measuring expectations and satisfaction with community-based psychiatric services in patients, relatives and professionals. *Psychol Med* 1993;**23**:511–23.

115. Han SH, Jung ES, Jung MY, Kwak JY, Park SJ, Choe JH. A psychophysical evaluation of interior-design alternatives for a high-speed train. *Comput Ind Eng* 1994;**27**:397–400.

116. Westra BL, Cullen L, Brody D, Jump P, Geanon L, Milad E. Development of the home care client satisfaction instrument. *Public Health Nurs* 1995;**12**:393–9.

117. Wilde B, Larsson G, Larsson M, Starrin B. Quality of care – development of a patient-centered questionnaire based on a grounded theory model. *Scand J Caring Sci* 1994;**8**:39–48.

118. Holmes-Rovner M, Kroll J, Schmitt N, Rovner DR, Breer ML, Rothert ML, *et al.* Patient satisfaction with health care decisions: the satisfaction with decision scale. *Med Decis Making* 1996;**16**:58–64.

119. Guyatt GH, Mitchell A, Molloy DW, Capretta R, Horsman J, Griffith L. Measuring patient and relative satisfaction with level or aggressiveness of care and involvement in care decisions in the context of life threatening illness. *J Clin Epidemiol* 1995;**48**:1215–24.

120. Fitzpatrick R. Surveys of patient satisfaction: II–Designing a questionnaire and conducting a survey. *BMJ* 1991;**302**:1129–32.

121. Campbell D, Christopher K. The evolution of a patient satisfaction survey. *Dimens Health Serv* 1991;**68**:18–20.

122. Van Campen C, Sixma H, Friele RD, Kerssens JJ, Peters L. Quality of care and patient satisfaction: a review of measuring instruments [review]. *Med Care Res Rev* 1995;**52**:109–33.

123. Bamford C, Jacoby A. Development of patient satisfaction questionnaires: 1. Methodological issues. *Qual Health Care* 1992;**1**:153–7.

124. Roberts RE, Pascoe GC, Attkisson CC. Relationship of service satisfaction to life satisfaction and perceived well-being. *Eval Program Plann* 1983;**6**:373–83.

125. Pascoe GC, Attkisson CC, Roberts RE. Comparison of indirect and direct approaches to measuring patient satisfaction. *Eval Program Plann* 1983;**6**:359–71.

126. Parker SC, Kroboth FJ. Practical problems of conducting patient-satisfaction surveys. *J Gen Intern Med* 1991;**6**:430–5.

127. McCusker J. Development of scales to measure satisfaction and preferences regarding long-term and terminal care. *Med Care* 1984;**22**:476–93.

128. Damkot DK, Pandiani JA, Gordon LR. Development, implementation, and findings of a continuing client satisfaction survey. *Community Mental Health J* 1983;**19**:265–78.

129. Hinshaw AS, Atwood JR. A Patient Satisfaction Instrument: precision by replication. *Nurs Res* 1982;**31**:170–5.

130. Ware JEJ. How to survey patient satisfaction. *Drug Intell Clin Pharm* 1981;**15**:892–9.

131. Hines BL, Clarkson QD, Smith DD. Development and use of a patient satisfaction questionnaire. *J Fam Pract* 1977;**4**:148–9.

132. Zyzanski SJ, Hulka BS, Cassel JC. Scale for the measurement of "satisfaction" with medical care: modifications in content, format and scoring. *Med Care* 1974;**12**:611–20.

133. Harrington V, Lackey NR, Gates MF. Needs of caregivers of clinic and hospice cancer patients. *Cancer Nurs* 1996;**19**:118–25.

134. Bredart A, Razavi D, Delvaux N, Goodman V, Farvacques C, VanHeer C. A comprehensive assessment of satisfaction with care for cancer patients. *Support Care Cancer* 1998;**6**:518–23.

135. Martin PD, Brantley PJ, McKnight GT, Jones GN, Springer A. The multidisciplinary hemodialysis patient satisfaction scale: reliability, validity, and scale development. *Assessment* 1997;**4**:95–105.

136. Singh J, Wood VR, Goolsby J. Consumers' satisfaction with health care delivery: issues of measurement, issues of research design [review]. *J Ambul Care Mark* 1990;**4**(1):105–15.

137. Cockburn J, Hill D, Irwig L, De Luise T, Turnbull D, Schofield P. Development and validation of an instrument to measure satisfaction of participants at breast screening programmes. *Eur J Cancer* 1991;**27**:827–31.

138. Ford RC, Bach SA, Fottler MD. Methods of measuring patient satisfaction in health care organizations [review]. *Health Care Manage Rev* 1997;**22**:74–89.

139. Batchelor C, Owens DJ, Read M. Patient satisfaction studies: methodology, management and consumer evaluation. *Int J Health Care Qual Assur* 1994;**7**(7):22–30.

140. Niedz BA. Correlates of hospitalized patients' perceptions of service quality. *Res Nurs Health* 1998;**21**:339–49.

141. Hugo B, Becker S, Witt E. Assessment of the combined orthodontic-surgical treatment from the patients' point of view. A longitudinal study. *J Orofacial Orthop* 1996;**57**:88–101.

142. Dansky KH, Brannon D. Discriminant analysis: a technique for adding value to patient satisfaction surveys. *Hosp Health Serv Adm* 1996;**41**:503–13.

143. Scott A, Smith RD. Keeping the customer satisfied: issues in the interpretation and use of patient satisfaction surveys. *Int J Qual Health Care* 1994;**6**:353–9.

144. Fitzpatrick R. Surveys of patient satisfaction: 1 – important general considerations. *BMJ* 1991;**302**:887–9.

145. Carr-Hill RA. The measurement of patient satisfaction. *J Public Health Med* 1992;**14**:236–49.

146. Baker R. The reliability and criterion validity of a measure of patients' satisfaction with their general practice. *Fam Pract* 1991;**8**:171–7.

147. Baker R. Development of a questionnaire to assess patients' satisfaction with consultations in general practice. *Br J Gen Pract* 1990;**40**:487–490.

148. McKinley RK, Manku-Scott T, Hastings AM, French DP, Baker R. Reliability and validity of a new measure of patient satisfaction with out of hours primary medical care in the United Kingdom: development of a patient questionnaire. *BMJ* 1997;**314**:193–8.

149. Kalman TP. An overview of patient satisfaction with psychiatric treatment. *Hosp Community Psychiatry* 1983;**34**:48–54.

150. Tarnowski KJ, Simonian SJ. Assessing treatment acceptance: the abbreviated acceptability rating profile. *J Behav Ther Exp Psychiatry* 1992;**23**:101–6.

151. Essex DW, Fox JA, Groom JM. The development, factor analysis, and revision of a client satisfaction form. *Community Mental Health J* 1981;**17**:226–35.

152. Hulka BS, Zyzanski SJ. Validation of a patient satisfaction scale: theory, methods and practice. *Med Care* 1982;**20**:649–53.

153. Nguyen TD, Attkisson CC, Stegner BL. Assessment of patient satisfaction: development of a service evaluation questionnaire. *Eval Program Plann* 1983;**6**:299–313.

154. Gonzalves PE, Minderler JJ, Tompkins DL. A patient satisfaction survey: a basis for changing delivery of services. *Mil Med* 1995;**160**:486–8.

155. Dull VT, Lansky D, Davis N. Evaluating a patient satisfaction survey for maximum benefit. *Jt Comm J Qual Improv* 1994;**20**:444–53.

156. Baker R, Whitfield M. Measuring patient satisfaction: a test of construct validity. *Qual Health Care* 1992;**1**:104–9.

157. Pope C, Mays N. Qualitative research: reaching the parts other methods cannot reach. An introduction to qualitative methods in health and health services research. *BMJ* 1995;**311**:42–5.

158. Sabourin S, Wright J, Duchesne C, Belisle S. Are consumers of modern fertility treatments satisfied? *Fertil Steril* 1991;**56**:1084–90.

159. Bisset A, Chesson R. Is this satisfaction survey satisfactory? Some points to consider in their planning and assessment. *Health Bull* 2000;**58**:45–52.

160. Green J. On the receiving end. *Health Serv J* 1988;**98**:880–2.

161. Wolf M, Putman S, James S, Stiles W. The Medical Interview Satisfaction Scale: development of a scale to measure patient perceptions of physician behaviour. *J Behav Med* 1978;**1**:391–401.

162. Eardley A, Lancaster G, Elkind A. Made to measure survey. *Health Serv J* 1990;**100**:1773.

163. Garcia J, Redshaw M, Fitzsimons B, Keene J. First Class Delivery: a national survey of women's views of maternity care. Oxford: Audit Commission; 1998.

164. Scottish Programme for Clinical Effectiveness in Reproductive Health, Dugald Baird Centre for Research On Womens' Health. Maternity care matters: an audit of maternity services in Scotland 1998. Aberdeen: SPCERH; 1999. SPCERH Publication No. 9.

165. Parasuraman A, Zeithaml VA, Berry LL. Servqual – a multiple-item scale for measuring consumer perceptions of service quality. *J Retailing* 1988;**64**:12–40.

166. Camilleri D, O'Callaghan M. Comparing public and private hospital care service quality. *Int J Health Care Qual Assur* 1998;**11**(4–5):127–33.

167. Scardina SA. SERVQUAL: a tool for evaluating patient satisfaction with nursing care. *J Nurs Care Qual* 1994;**8**(2):38–46.

168. Sargeant A, Kaehler J. Factors of patient satisfaction with medical services: the case of GP practices in the UK. *Health Mark Q* 1998;**16**(1):55–77.

169. Kaldenberg D, Becker BW, Browne BA, Browne WG. Identifying service quality strengths and weaknesses using SERVQUAL: a study of dental services. *Health Mark Q* 1997;**15**(2):69–86.

170. Sewell N. Continuous quality improvement in acute health care: creating a holistic and integrated approach. *Int J Health Care Qual Assur* 1997;**10**(1):20–6.

171. Youssef FN, Nel D, Bovaird T. Health care quality in NHS hospitals. *Int J Health Care Qual Assur* 1996;**9**(1):15–28.

172. Anderson EA. Measuring service quality at a university health clinic. *Int J Health Care Qual Assur* 1995;**8**(2):32–7.

173. Headley DE, Miller SJ. Measuring service quality and its relationship to future consumer behavior. *J Health Care Mark* 1993;**13**(4):32–41.

174. Walbridge SW, Delene LM. Measuring physician attitudes of service quality. *J Health Care Mark* 1993;**13**(1):6–15.

175. Soliman AA. Assessing the quality of health care: a consumerist approach. *Health Mark Q* 1992;**10**(1–2):121–41.

176. Reidenbach RE, Sandifer-Smallwood B. Exploring perceptions of hospital operations by a modified SERVQUAL approach. *J Health Care Mark* 1990;**10**(4):47–55.

177. Lam SSK. SERVQUAL: a tool for measuring patients' opinions of hospital service quality in Hong Kong. *Total Qual Manage* 1997;**8**:145–52.

178. Duffy JA, Duffy M, Kilbourne W. Cross national study of perceived service quality in long-term care facilities. *J Aging Stud* 1997;**11**:327–36.

179. Chaston I. A comparative-study of internal customer management practices within service sector firms and the National Health Service. *J Adv Nurs* 1994;**19**:299–308.

180. Mitchell R, Leanna JC, Hyde R. Client satisfaction with nursing services: evaluation in an occupational health setting. *AAOHN J* 1999;**47**(2):74–8.

181. Raspollini E, Pappalettera M, Riccardi D, Parravicini A, Sestili S, Rebulla P, *et al.* Use of SERVQUAL to assess clinicians' satisfaction with the blood transfusion service. *Vox Sang* 1997;**73**:162–6.

182. Duffy J-AM, Ketchand AA. Examining the role of service quality in overall service satisfaction. *J Manage Issues* 1998;**10**(2):240–55.

183. Charny MC, Lewis PA, Farrow SC. Choosing who shall not be treated in the NHS. *Soc Sci Med* 1989;**28**:1331–8.

184. Lewis PA, Charny M. Which of two individuals do you treat when only their ages are different and you can't treat both? *J Med Ethics* 1999;**15**:28–32.

185. Mooney G, Jan S, Wiseman V. Examining preferences for allocating health care gains. *Health Care Anal* 1995;**3**:261–5.

186. Ryynanen O-P, Myllykangas M, Vaskilampi T, Takala J. Random paired scenarios – a method for investigating attitudes to prioritisation in medicine. *J Med Ethics* 1996;**22**:238–42.

187. Bryan S, Buxton M, Sheldon R, Grant A. Magnetic resonance imaging for the investigation of knee injuries: an investigation of preferences. *Health Econ* 1998;**7**:595–603.

188. Chakraborty G, Ettenson R, Gaeth G. How consumers choose health insurance. *J Health Care Mark* 1994;**14**(1):21–33.

189. Farrar S, Ryan M. Response-ordering effects: a methodological issue in conjoint analysis. *Health Econ* 1999;**8**:75–9.

190. Ryan M. A role for conjoint analysis in technology assessment in health care. *Int J Technol Assess Health Care* 1999;**15**:443–57.

191. Ferguson RP, Wetle T, Dubitzky D, Winsemius D. Relative importance to elderly patients of effectiveness, adverse effects, convenience and cost of antihyperintensive medications. *Drugs Aging* 1994;**4**:56–62.

192. Hakim Z, Pathak DS. Modelling the EuroQol data: a comparison of discrete choice conjoint and conditional preference modelling. *Health Econ* 1999;**8**:103–16.

193. McIntosh E, Ryan M. The theoretical basis of discrete choice conjoint analysis: testing for transitivity and continuity within an empirical application in health care. Paper presented to the UK Health Economists' Study Group Meeting; 1999 Jul; Aberdeen.

194. Ryan M. Measuring benefits in health care: the role of discrete choice conjoint analysis. Paper presented to the 2nd International Conference of the International Health Economics' Association; 1999 Jun 7–9; Rotterdam.

195. Propper C. Contingent valuation of time spent on NHS waiting lists. *Econ J* 1990;**100**:193–9.

196. Ratcliffe J, Buxton M. Patients' preferences regarding the process and outcomes of life. *Int J Technol Assess Health Care* 1999;**15**:340–51.

197. Ryan M. Using conjoint analysis to take account of patient preferences and go beyond health outcomes: an application to in vitro fertilisation. *Soc Sci Med* 1999;**48**:535–46.

198. Ryan M. A role for conjoint analysis in technology assessment in health care? *Int J Technol Assess Health Care* 1999;**15**:443–57.

199. Ryan M, Hughes J. Using conjoint analysis to assess women's preferences for miscarriage management. *Health Econ* 1997;**6**:261–73.

200. Ryan M, Wordsworth S. Sensitivity of willingness to pay estimates to the level of attributes in discrete choice experiments. *Scott J Polit Econ* 2000;**47**:504–24.

201. San Miguel F, Ryan M, McIntosh E. Demonstrating the use of conjoint analysis in economic evaluations: an application to menorrhagia. *Appl Econ* 2000;**32**:823–33.

202. Scott A, Vick S. An application of principle–agent theory to the doctor–patient relationship. *Scott J Polit Econ* 1999;**46**:111–34.

203. Vick S, Scott A. Agency in health care. Examining patients' preferences for attributes. *J Health Econ* 1998;**17**:587–605.

204. Hadorn DC, Hays RD, Uebersax J, Hauber T. Improving task comprehension in the measurement of health state preferences. A trial of informational cartoon figures and a paired-comparison task. *J Clin Epidemiol* 1992;**45**:233–43.

205. Ryan M, Farrar S. Eliciting preferences for health care using conjoint analysis. *BMJ* 2000;**320**:1530–3.

206. Spoth R, Redmond C. Identifying program preferences through conjoint analysis: illustrative results from a parent sample. *Am J Health Promot* 1993;**8**:124–33.

207. Van der Pol M, Cairns J. Establishing patient preferences for blood transfusion support: an application of conjoint analysis. *J Health Serv Res Policy* 1998;**3**:70–6.

208. Spoth R, Redmond C, Ball A. Stage of quitting and motivational factors relevant to smoking cessation program choices. *Psychol Addict Behav* 1993;**7**:29–42.

209. Spoth R. Multi-attribute analysis of benefit managers' preferences for smoking cessation programs. *Health Values* 1990;**14**(5):3–15.

210. Hornberger JC, Habraken H, Bloch DA. Minimum data needed on patient preferences for accurate, efficient medical decision making. *Med Care* 1995;**33**:297–310.

211. Farrar S, Ryan M, Ross D, Ludbrook A. Using discrete choice modelling in priority setting: an application to clinical service developments. *Soc Sci Med* 2000;**50**:63–75.

212. Verhoef CG, Maas A, Stalpers LJA, Verbeek ALM, Wobbes T, van Daal WAJ. The feasibility of additive conjoint measurement in measuring utilities in breast cancer patients. *Health Policy* 1991;**17**:39–50.

213. Maas A, Stalpers L. Assessing utilities by means of conjoint measurement: an application in medical decision analysis. *Med Decis Making* 1992;**12**:288–97.

214. Rangone A. An analytical hierarchy process framework for comparing the overall performance of manufacturing departments. *Int J Oper Prod Manage* 1996;**16**(8):104–19.

215. Saaty RW. The analytic hierarchy process – what it is and how it is used. *Math Model* 1987;**9**:161–76.

216. Mulye R. An empirical comparison of three variants of the AHP and two variants of conjoint analysis. *J Behav Decis Making* 1998;**11**:263–80.

217. Golden BL, Wasil EA, Harker PT. The analytic hierarchy process: applications and studies. New York: Springer-Verlag; 1989.

218. Harker PT, Vargas LG. The theory of ratio scale estimation – saaty's analytic hierarchy process. *Management Sci* 1987;**33**:1383–403.

219. Saaty TL. Axiomatic foundation of the analytic hierarchy process. *Management Sci* 1986;**32**:841–55.

220. Hanratty PJ, Joseph B. Decision-making in chemical-engineering and expert systems – application of the analytic hierarchy process to reactor selection. *Comput Chem Eng* 1992;**16**:849–60.

221. Mohanty RP, Venkataraman S. Use of analytic hierarchy process for selecting automated manufacturing systems. *Int J Oper Prod Manage* 1993;**13**(8):45–57.

222. Olson DL, Venkataramanan M, Mote JL. A technique using analytical hierarchy process in multiobjective planning-models. *Socioecon Plann Sci* 1986;**20**:361–8.

223. Saaty TL. A scaling method for priorities in hierarchical structures. *J Math Psychol* 1977;**15**:234–81.

224. Mohanty RP, Deshmukh SG. Use of analytic hierarchy process for evaluating sources of supply. *Int J Phys Distribution Logistics Manage* 1993;**23**(3):22–8.

225. Schwartz RG, Oren S. Using analytic hierarchies for consumer research and market modelling. *Math Comput Model* 1988;**11**:266–71.

226. Dolan JG. Choosing initial antibiotic therapy for acute pyleonephritis. In: Golden BLWEA, Harker PT, editors. The analytic hierarchy process: applications and studies. London: Springer-Verlag; 1989. p. 213–24.

227. Dolan JG, Isselhardt BJ, Cappuccio JD. The analytic hierarchy process in medical decision making: a tutorial. *Med Decis Making* 1989;**9**:40–50.

228. Dolan JG. Medical decision making using the analytic hierarchy process: choice of initial antimicrobial therapy for acute pyelonephritis. *Med Decis Making* 1999;**9**:51–6.

229. Dolan JG, Bordley DR. Isoniazid prophylaxis: the importance of individual values. *Med Decis Making* 1994;**14**:1–8.

230. Dolan JG, Bordley DR. Should concern over gastric cancer influence the choice of diagnostic tests in patients with acute upper gastrointestinal bleeding. Proceedings of the 2nd International Symposium on the Analytic Hierarchy Process; 1991 Aug 12–14; Pittsburgh (PA). p. 391–403.

231. Dolan JG. Promoting decision making partnerships using the Analytic Hierarchy Process. In: Suchman AL, Botelho RJ, Hinton-Walker P, editors. Partnerships in health care: transforming relational process. Rochester, NY: University of Rochester Press; 1998. p. 151–66.

232. Dolan JG. Diagnostic strategies in the management of acute upper gastrointestinal bleeding: patient and physician preferences. *J Gen Intern Med* 1993;**8**:525–9.

233. Dolan JG, Bordley DR. Using the Analytic hierarchy Process (AHP) to develop and disseminate guidelines. *Q Rev Bull* 1992;**18**:440–7.

234. Dolan JG. Clinical decision making using the analytic hierachy process: choice of antibiotic treatment for community-acquired pyelonephritis [abstract]. *Clin Res* 1987;**35**:738.

235. Dolan JG. Are patients capable of using the analytic hierarchy process and willing to use it to help make clinical decisions. *Med Decis Making* 1995;**15**:76–80.

236. Dolan JG. Involving patients in complex decisions about their care: an approach using the analytic hierarchy process. *J Gen Intern Med* 1993;**8**:204–9.

237. Dolan JG. Can decision analysis adequately represent clinical problems? *J Clin Epidemiol* 1990;**43**:277–84.

238. Lusk EJ. Analysis of hospital capital decision alternatives: a priority assessment model. *J Oper Res Soc* 1979;**30**:439–48.

239. Javalgi RG, Rao SR, Thomas EG. Choosing a hospital: analysis of consumer tradeoffs. *J Health Care Mark* 1989;**11**(1):12–22.

240. Saaty TL. The analytic hierarchy process and health care problems. In: Tilquin C, editor. Systems science in health care. Vol. 1. Proceedings of the International Conference on Systems Science in Health Care; 1980 Jul; Montreal. Toronto: Pergamon Press; 1981. p. 147–58.

241. Hannan EL, O'Donnell J, Freedland T. A priority assessment model for standards and conditions in a long term care survey. *Socioecon Plann Sci* 1981;**15**:277–89.

242. Dougherty JJ, Saaty TL. Optimum determination of hospital requirements. In: Saaty TL, Vargas LG, editors. The logic of priorities: applications in business, energy, health and transportation. Boston: Kluwer-Nijhoff Publishing; 1982. p. 165–81.

243. Odynocki B. The NHI proposals: an evaluation. In: Tilquin C, editor. Systems science in health care. Vol. 1. Proceedings of the International Conference on Systems Science in Health Care; 1980 Jul; Montreal. Toronto: Pergamon Press; 1981. p. 159–68.

244. Odynocki B. The forward and backward processes in health policy planning. *Math Comput Simulation* 1983;**25**:146–55.

245. Gafni A. The standard gamble method: what is being measured and how it is interpreted. *Health Services Res* 1994;**29**:207–24.

246. Bleichrodt H, Johannesson M. Standard gamble, time trade-off and rating scale: experimental results on the ranking properties of QALYs. *J Health Econ* 1997;**16**:155–75.

247. Torrance G, Thomas W, Sackett D. A utility maximisation model for evaluation of health care programs. *Health Serv Res* 1972;**7**:118–33.

248. Patrick DL, Bush JW, Chen MM. Methods for measuring levels of well-being for a health status index. *Health Serv Res* 1973;**8**:228–45.

249. Nord E. Methods for quality adjustment of life years [review]. *Soc Sci Med* 1992;**34**:559–69.

250. Green C. Assessing the societal value of health care: is the PTO up to the job? Paper presented to the UK Health Economists' Study Group Meeting; 1999 Jul; Aberdeen.

251. Nord E. The person-trade-off approach to valuing health care programs. *Med Decis Making* 1995;**15**:201–8.

252. Prades JLP. Is the person trade-off a valid method for allocating health care resources? *Health Econ* 1997;**6**:71–81.

253. Brazier J, Deverill M, Green C, Harper R, Booth A. A review of the use of health status measures in economic evaluation. *Health Technol Assess* 1999;**3**(9).

254. Ubel PA, Loewenstein G, Scanlon D, Kamlet M. Value measurement in cost-utility analysis: explaining the discrepancy between rating scale and person trade-off elicitations. *Health Policy* 1998;**43**:33–44.

255. Nord E. The person trade-off approach to valuing health care programs. Fairfield, Vic: National Centre for Health Program Evaluation, Monash University/University of Melbourne; 1994. CHPE Working Paper 38.

256. Diener A, Obrien B, Gafni A. Health care contingent valuation studies: a review and classification of the literature. *Health Econ* 1998;**7**:313–26.

257. Klose T. The contingent valuation method in health care. *Health Policy* 1999;**47**:97–123.

258. Olsen JA, Smith R. Who have been asked to value what? A review of 54 WTP surveys on health and health care. Paper presented to the UK Health Economists' Study Group Meeting; 1998 Jul; Galway.

259. Donaldson C, Jones AM, Mapp TJ, Olson JA. Limited dependent variables in willingness to pay studies: applications in health care. *Appl Econ* 1998;**30**:667–77.

260. Ryan M, Ratcliffe J, Tucker J. Using willingness to pay to value alternative models of antenatal care. *Soc Sci Med* 1997;**44**:371–80.

261. Johannesson M, Jonsson B, Borgquist L. Willingness to pay for antihyperintensive therapy – results of a Swedish pilot study. *J Health Econ* 1991;**10**:461–74.

262. Olsen JA, Donaldson C. Helicopters, hearts and hips: using willingness to pay to set priorities for public sector health care programmes. *Soc Sci Med* 1998;**46**:1–12.

263. Johannesson M, Johansson P-O, Kristrom B, Gerdtham U-G. Willingness to pay for antihyper-intensive therapy – further results. *J Health Econ* 1993;**12**:95–108.

264. Churchman CW, Ackoff RL. An approximate measure of value. *Oper Res* 1954;**2**:172–81.

265. Dickinson AL. Getting the best from additional resources. *Hosp Health Serv Rev* 1979;**75**:127–9.

266. Secker J. Qualitative methods in health promotion research: some criteria for quality. *Health Educ J* 1995;**54**:74–87.

267. Hoinville G. The Priority Evaluator Method. London: Department of Social and Community Planning and Research, University of London; 1977. Methodological Working Paper 3.

268. Hoinville G. Evaluating community preferences. *J Mark Res Soc* 1996;**38**:483–504.

269. Ruta DA, Garratt AM, Leng M, Russell IT, MacDonald LM. A new approach to the measurement of quality of life – the patient generated index. *Med Care* 1999;**32**:1109–26.

270. Macduff C, Russell E. The problem of measuring change in individual health-related quality of life by postal questionnaire: use of the patient-generated index in a disabled population. *Qual Life Res* 1998;**7**:761–9.

271. Ruta DA, Garratt AM, Russell IT. Patient centred assessment of quality of life for patients with four common conditions. *Qual Health Care* 1999;**8**:22–9.

272. Herd RM, Tidman MJ, Ruta DA, Hunter JA. Measurement of quality of life in atopic dermatitis: correlation and validation of two different methods. *Br J Dermatol* 1997;**136**:502–7.

273. Jenkinson C, Stradling J, Petersen S. How should we evaluate health status? A comparison of three methods in patients presenting with obstructive sleep apnoea. *Qual Life Res* 1998;**7**:95–100.

274. Hickey AM, Bury G, O'Boyle CA, Bradley F, O'Kelly FD, Shannon W. A new short form individual quality of life measure (SEIQoL-DW): application in a cohort of individuals with HIV/AIDS. *BMJ* 1996;**313**:29–33.

275. Browne JP, O'Boyle C, McGee HM, McDonald NJ, Joyce CRB. Development of a direct weighting procedure for quality of life domains. *Qual Life Res* 1997;**6**:301–9.

276. Clark TN. Community social indicators: from analytical models to policy applications. *Urban Affairs Q* 1973;**9**:3–36.

277. Clark TM. Can you cut a budget pie? *Policy Politics* 1974;**3**(2):3–31.

278. Hauser JR, Shugan SM. Intensity measures of consumer preference. *Oper Res* 1980;**28**:278–320.

279. Ham C, Locock L. International approaches to priority setting in health care: an annotated listing of official and semi-official publications, with a selection of key academic references. Birmingham: Health Services Management Centre, University of Birmingham; 1998. HSMC Handbook Series no. 25.

280. Honigsbaum F, Richards J, Lockett T. Priority setting in action: purchasing dilemmas. Oxford: Radcliffe Medical Press; 1995.

281. Ratcliffe J. Distribution principles for the allocation of donor liver grafts: results from a 'social' conjoint analysis survey. Paper presented at the UK Health Economists' Study Group Meeting; 1999 Jan; Birmingham.

282. Strauss RP, Hughes GD. A new approach to the demand for public goods. *J Public Econ* 1976;**6**:191–204.

283. Garland MJ. Rationing in public: Oregon's priority-setting methodology. In: Strosberg MA, editor. Rationing America's medical care: the Oregon plan and beyond. Washington, DC: The Brookings Institution; 1992. p. 37–59.

284. Dixon J, Welch HG. Priority setting: lessons from Oregon. *Lancet* 1991;**337**:891–4.

285. Eddy DM. What's going on in Oregon? *JAMA* 1991;**266**:417–20.

286. Klein R. On the Oregon trail: rationing health care [editorial] *BMJ* 1991;**302**:1–2.

287. Foster G. Fishing with the net for research data. *Br J Educ Technol* 1994;**25**:91–7.

288. Ayanian JZ, Cleary PD, Weissman JS, Epstein AM. The effect of patients' preferences on racial differences in access to renal transplantation. *N Engl J Med* 1999;**341**:1661–9.

289. Dicker A, Armstrong D. Patients' views of priority setting in health care: an interview survey in one practice. *BMJ* 1995;**311**:1137–9.

290. Williams B, Coyle J, Healy D. The meaning of patient satisfaction: an explanation of high reported levels. *Soc Sci Med* 1998;**47**:1351–9.

291. Crabtree B, Miller WL. A qualitative approach to primary care research: the long interview. *Fam Med* 1991;**23**:145–51.

292. 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.

293. Sohier R. The dyadic interview as a tool for nursing research. *Appl Nurs Res* 1995;**8**:96–101.

294. Guest JF, Hart WM, Cookson RF. Cost analysis of palliative care for terminally ill cancer patients in the UK after switching from weak to strong opoids. *Pharmacoeconomics* 1998;**14**:241–339.

295. Endacott R, Clifford CM, Tripp JH. Can the needs of the critically ill child be identified using scenarios? Experiences of a modified Delphi study. *J Adv Nurs* 1999;**30**:665–76.

296. Harrington JM. Research priorities in occupational medicine: a survey of United Kingdom medical opinion by the Delphi technique. *Occup Environ Med* 1993;**51**:289–94.

297. Charlton JRH, Patrick DL, Matthews G, West PA. Spending priorities in Kent: a Delphi study. *J Epidemiol Community Health* 1981;**35**:288–92.

298. Roberts H, Khee TS, Philip I. Setting priorities for measures of performance for geriatric medical services. *Age Ageing* 1994;**23**:154–7.

299. Hadorn DC, Holmes AC. The New Zealand priority criteria project. Part 1: Overview. *BMJ* 1997;**314**:131–4.

300. Gabbay J, Francis L. How much day surgery? Delphic predictions. *BMJ* 1988;**297**:1249–52.

301. Wilson DM, Kerr JR. An exploration of Canadian social values relevant to health care. *Am J Health Behav* 1998;**22**:120–9.

302. Thomson WA, Ponder LD. Use of Delphi methodology to generate a survey instrument to identify priorities for state allied health associations. *Allied Health Behav Sci* 1979;**2**:383–99.

303. Gallagher M, Bradshaw C, Nattress H. Policy priorities in diabetes care: a Delphi study. *Qual Health Care* 1996;**5**:3–8.

304. Burns TJ, Batavia AI, Smith QW, DeJong G. Primary health care needs of persons with physical disabilities: what are the research and service priorities? *Arch Phys Med Rehabil* 1990;**71**:138–43.

305. Kastein MR, Jacobs M, Roelof HvdH, Luttik K, Touw-Otten F. Delphi, the issue of reliability: a qualitative Delphi study in primary health care in the Netherlands. *Technological Forecasting Soc Change* 1993;**44**:315–23.

306. Festervand TA. An introduction and application of focus group research to the health care industry. *Health Mark Q* 1984;**2**(2–3):199–209.

307. Morgan DL. Focus groups as qualitative research. 2nd ed. Thousand Oaks, CA: Sage; 1997.

308. Dowell J, Huby G, Smith C. Scottish Consensus Statement on quality research in primary health care. Dundee: Tayside Centre for General Practice, University of Dundee; 1995.

309. Morgan D. Focus groups as qualitative research. Newbury Park, CA: Sage; 1988.

310. Krueger RA. Focus groups: a practical guide for applied research. Newbury Park, CA: Sage; 1988.

311. Bowie C, Richardson A, Sykes W. Consulting the public about health service priorities. *BMJ* 1995;**311**:1155–8.

312. Kuder LB, Roeder PW. Attitudes toward age-based health care rationing: a qualitative assessment. *J Aging Health* 1995;**7**:301–27.

313. Cohen MB, Garrett KJ. Breaking the rules: a group work perspective on focus group research. *Br J Soc Work* 1999;**29**:359–72.

314. Kitzinger J. The methodology of focus groups: the importance of interaction between research participants. *Sociol Health Illn* 1994;**16**:103–21.

315. Ward VM, Bertrand JT, Brown LF. The comparability of focus group and survey results: three case studies. *Eval Rev* 1991;**15**:266–83.

316. Carey MA, Smith MW. Enhancement of validity through qualitative approaches – incorporating the patients perspective. *Eval Health Prof* 1992;**15**:107–14.

317. Powell RA, Single HM, Lloyd KR. Focus groups in mental health research: enhancing the validity of user and provider questionnaires. *Int J Soc Psychiatry* 1996;**42**:193–206.

318. Wilkinson S. Focus groups in health research. *J Health Psychol* 1998;**3**:329–48.

319. Stevens PE. Focus groups: collecting aggregate-level data to understand community health phenomena. *Public Health Nurs* 1996;**13**:170–6.

320. Keller KL, Sliepcevich EM, Vitello EM, Lacey EP, Wright WR. Assessing beliefs about and needs of senior citizens using the focus group interview: a qualitative approach. *Health Educ* 1987;**18**:44–9.

321. Smith B, West P. Focus groups: giving voice to the community. *Perspectives* 1998;**22**(3):2–7.

322. Bradley N, Sweeney K, Waterfield M. The health of their nation: how would citizens develop England's health strategy? *Br J Gen Pract* 1999;**49**:801–5.

323. Dolan P, Cookson R, Ferguson B. Effect of discussion and deliberation on the public's views of priority setting in health care: focus group study. *BMJ* 1999;**318**:916–19.

324. Ramirez AG, Shepperd J. The use of focus groups in health research. *Scand J Prim Health Care* 1988; **6** Suppl 1:S81–S90.

325. Weinberger M, Ferguson JA, Westmoreland G, Mamlin LA, Segar DS, Eckert GJ, *et al.* Can raters consistently evaluate the content of focus groups. *Soc Sci Med* 1998;**46**:929–33.

326. Southern DM, Batterham RW, Appleby NJ, Young D, Dunt D, Guibert R. The concept mapping method: an alternative to focus group inquiry in general practice. *Aust Fam Physician* 1999;**28** Suppl 1:S35–S40.

327. Trochim WMK, Linton R. Conceptualization for planning and evaluation. *Eval Program Plann* 1986;**9**:289–308.

328. Lenaghan J. Involving the public in rationing decisions. The experience of citizens juries. *Health Policy* 1999;**49**:45–61.

329. McIver S. Healthy debate? An independent evaluation of citizens' juries in health settings. London: King's Fund; 1998.

330. Renn O, Webler T, Rakel H, Dienel P, Johnson B. Public-participation in decision-making – a 3-step procedure. *Policy Sci* 1993;**26**:189–214.

331. Coote A, Lenaghan J. Citizens' juries: theory into practice. London: Institute for Public Policy Research; 1997.

332. Lenaghan J, New B, Mitchell E. Setting priorities: is there a role for citizens' juries? *BMJ* 1996;**312**:1591–3.

333. Davies S, Elizabeth S, Hanley S, New B, Sang B. Ordinary wisdom. London: King's Fund; 1998.

334. Dunkerley D, Glasner P. Empowering the public? Citizens' juries and the new genetic technologies. *Crit Public Health* 1998;**8**:181–92.

335. Reid N, Reid R, Morris D. Customer complaints in the National Health Service. *J Nurs Manag* 1995;**3**:295–9.

336. Stronks K, Strijbis AM, Wendte JF. Who should decide? Qualitative analysis of panel data from public, patients, healthcare professionals, and insurers on priorities in health care. *BMJ* 1997;**315**:92–6.

337. Rosenthal TC, Ferrara E, Hesler E. Providing birthing services in rural health networks: coping with change in New York State. *J Rural Health* 1996;**12**:137–45.

338. Hirschfield RM, Keller MB, Panico S, Arons BS, Barlow D, Dandoff F, *et al.* The National Depressive and Manic-Depressive Association Consensus Statement on the Undertreatment of Depression. *JAMA* 1997;**277**:333–40.

339. Fernandez AM, Schrogie JJ, Wilson WW, Nash D. Technology assessment in healthcare: a review and description of a "best practice" technology assessment process. *Best Pract Benchmarking Healthc* 1997;**2**:240–53.

340. Coulter I, Adams A, Shekelle P. Impact of varying panel membership on ratings of appropriateness in consensus panels: a comparison of a multi- and single disciplinary panel. *Health Serv Res* 1995;**30**:577–91.

341. Gundry KG, Heberlein TA. Do public meetings represent the public? *J Am Planning Assoc* 1984;**50**:175–82.

342. Broadbent B. Open to question. *Health Serv J* 1998;**108**:30–1.

343. Gott M, Warren G. Neighbourhood health forums: local democracy at work. *World Health Forum* 1991;**12**:413–18.

344. Gallagher M, Hares T, Spencer J, Bradshaw C, Webb I. The nominal group technique: a research tool for general practice? *Fam Pract* 1993;**10**:76–81.

345. DeBold B. The nominal group technique [Internet communication], 2000; http://www.radix.net/~ash2jam/TQM/nominal.htm

346. Rohrbaugh J. Improving the quality of group judgment: social judgment analysis and the nominal group technique. *Organ Behav Hum Perform* 1981;**28**:272–88.

347. Hares T, Spencer J, Gallagher M, Bradshaw C, Webb I. Diabetes care: who are the experts? *Qual Health Care* 1992;**1**:219–24.

348. Jones J, Hunter D. Concensus methods for medical and health services research. *BMJ* 1995;**311**:376–80.

349. Redman S, Carrick S, Cockburn J, Hirst S. Consulting about priorities for the NHMRC National Breast Cancer Centre: how good is the nominal group technique. *Aust N Z J Public Health* 1997;**21**:250–6.

350. Streiner DL, Norman GR. Health Measurement Scales: a practical guide to their development and use. Oxford: Oxford University Press; 1989.

351. Crombie IK, Davies.HTO. Research in health care: design, conduct and interpretation of health services research. Chichester: John Wiley; 1996.

352. Froberg DG, Kane RL. Methodology for measuring health-state preferences–II: scaling methods. *J Clin Epidemiol* 1989;**42**:459–71.

353. Phillips KA, Homan RK, Luft HS, Hiatt PH, Olson KR, Kearney TE, *et al.* Willingness to pay for poison control centers. *J Health Econ* 1997;**16**:343–57.

354. Donaldson C. Developing the method of 'willingness to pay' for assessment of community preferences for health care. Final report to Biomed 2 Programme (PL950832) of the European Commission. Health Economics Research Unit, University of Aberdeen and Departments of Economics and Community Health Sciences, University of Calgary, 1999.

355. Donaldson C, Shackley P, Abdalla M. Using willingness to pay to value close substitutes: carrier screening for cystic fibrosis revisited. *Health Econ* 1997;**6**:145–59.

356. Chestnut L, Keller L, Lambert W, Rowe R. Measuring heart patients' willingness to pay for changes in angina symptoms. *Med Decis Making* 1996;**16**:65–77.

357. San Miguel F, Ryan M, Scott A. Testing the assumptions of completeness and stability of preferences in discrete choice experiments. Paper presented at the 2nd International Conference of the International Health Economics Association; 1999 Jun; Rotterdam.

358. Shickle D. Public preferences for health care: prioritisation in the United Kingdom. *Bioethics* 1997;**11**:277–90.

359. Gudex C, Dolan P, Kind P, Williams A. Health state valuations from the general public using the visual analogue scale. *Qual Life Res* 1996;**5**:521–31.

360. Dolan P, Kind P. Inconsistency and health state valuations. *Soc Sci Med* 1996;**42**:609–15.

361. Llewellyn-Thomas H, Sutherland HJ, Tibshirani R, Ciampi A, Till JE, Boyd NF. The measurement of patients' values in medicine. *Med Decis Making* 1982;**2**:449–62.

362. Dolan P, Gudex C, Kind P, Williams A. The Time Trade-Off Method: results from a general population study. *Health Econ* 1996;**5**:141–54.

363. Ashby J, O'Hanlon M, Buxton MJ. The time trade-off technique: how do the valuations of breast cancer patients compare to those of other groups? *Qual Life Res* 1994;**3**:257–65.

364. Ubel PA, Loewenstein G, Scanlon D, Kamlet M. Individual utilities are inconsistent with rationing choices: a partial explanation of why Oregon's cost-effectiveness list failed. *Med Decis Making* 1996;**16**:108–16.

365. Badia X, Roset M, Herdman M. Inconsistent responses in three preference-elicitation methods for health states. *Soc Sci Med* 1999;**49**:943–50.

366. Bowling A. What people say about prioritising health services. London: King's Fund; 1999.

367. Frew E, Wolstenholme J, Whynes D. Willingness to pay for colorectal cancer screening: faecal occult blood test versus flexible sigmoidoscopy. Paper presented to the UK Health Economists' Study Group Meeting; 1999 Jul; Aberdeen.

368. Miedzybrodzka Z, Semper J, Shackley P, Abdalla M, Donaldson C. Stepwise or couple antenatal carrier screening for cystic fibrosis? Women's preferences and willingness to pay. *J Med Genet* 1995;**32**:282–3.

369. Ryan M, San Miguel F. Testing for consistency in willingness to pay experiments. *J Econ Psychol* 2000;**21**:305–17.

370. Babakus E, Mangold WG. Adapting the SERVQUAL scale to hospital services: an empirical investigation. *Health Serv Res* 1992;**26**:767–86.

371. Murphy E, Dingwall R, Greatbatch D, Parker S, Watson P. Qualitative research methods in health technology assessment: a review of the literature. *Health Technol Assess* 1998;**2**(14).

372. Mason J. Qualitative researching. London: Sage; 1996.

373. Dingwall R, Murphy E, Watson P, Greatbatch D, Parker S. Catching goldfish: quality in qualitative research. *J Health Serv Res Policy* 1998;**3**:167–72.

374. Lincoln YS, Guba E. Naturalistic enquiry. Beverly Hills, CA: Sage; 1985.

375. Seale C. The quality of qualitative research. London: Sage; 1999.

376. Kumar NC, Ganesh LS. A simulation-based evaluation of the approximate and the exact eigenvector methods employed in AHP. *Eur J Oper Res* 1996;**95**:656–62.

377. Blaikie NWH. A critique of the use of triangulation in social research. *Qual Quantity* 1991;**25**:115–36.

378. Silverman D, editor. Qualitative research: theory, method and practice. London: Sage; 1997.

379. Walker R, editor. Applied qualitative research. Aldershot: Gower; 1985.

380. MacPherson IA, Williamson PJ. 'Not quite what I meant!' – techniques of respondent validation. *Res Policy Plann* 1992;**10**:10–13.

381. Banister P, Burman E, Parker I, Taylor M, Tindall C. Qualitative methods in psychology: a research guide. Buckingham: Open University Press; 1994.

382. Bryman A, Burgess R. Analyzing qualitative data. London: Routledge; 1994.

383. Conrad P, Reinharz S. Computers and qualitative data. Editor's introductory essay. *Qual Sociol* 1984;**7**:1–2.

384. Glaser BG, Strauss AL. The discovery of grounded theory: strategies for qualitative research. Chicago: Aldine Publishing Co.; 1967.

385. Strauss AL. Qualitative analysis for social scientists. Cambridge: Cambridge University Press; 1987.

386. Denzin NK, Lincoln YS, editors. Handbook of qualitative research. Thousand Oaks, CA: Sage; 1994.

387. Fink A. Evaluation fundamentals: guiding health programs, research, and policy. Newbury Park, CA: Sage; 1993.

388. Fitzpatrick R, Davey C, Buxton M, Jones DR. Evaluating patient-based outcome measures for use in clinical trials. *Health Technol Assess* 1988;**2**(14).

389. Kirk J, Miller M. Reliability and validity in qualitative research. Newbury Park, CA: Sage; 1986.

390. O'Brien B, Gafni A. When do the "dollars" make sense? Toward a conceptual framework for contingent valuation studies in health care. *Med Decis Making* 1996;**16**:299.

391. Morrison GC, Gyldmark M. Appraising the use of contingent valuation. *Health Econ* 1992;**1**:233–43.

392. Arrow K, Solow R, Portney P, Leamer E, Radner R, Schuman H. Report of the NOAA panel of contingent valuation. *Federal Register* 1993;**10**:4601–14.

393. Drummond MF, Stoddart GL, Torrance GW. Methods for the economics evaluation of health care programmes. 2nd ed. Oxford: Oxford University Press; 1997.

394. CASP's qualitative checklist. 1999. URL: http://www.phru.org/casp/qualitative.html

395. Mays N, Pope C. Rigour and qualitative research. *BMJ* 1995;**311**:109–12.

396. Shackley P, Ryan M. Involving consumers in health care decision making. *Health Care Anal* 1995;**3**:196–204.

397. Deleted.

398. Green PCF, Wind Y. Subjective evaluation models and conjoint measurement. *Behav Sci* 1972;**17**:288–99.

399. Nord E. The validity of a visual analogue scale in determining social utility weights for health states. *Int J Health Plann Manage* 1991;**6**:234–42.

400. Torrance GW. Social preferences for health states: an empirical evaluation of three measurement techniques. *Socioecon Plann Sci* 1976;**10**:129–36.

401. Kaplan RM, Bush JW, Berry CC. Health status index: category rating versus magnitude estimation for measuring levels of well-being. *Med Care* 1979;**17**:501–25.

402. Sintonen H. An approach to measuring and valuing health states. *Soc Sci Med* 1981;**15**:55–65.

403. Richardson J. Cost utility analysis: what should be measured? *Soc Sci Med* 1994;**39**:7–21.

404. Johannesson M, Jonsson B, Karlsson G. Outcome measurement in economic evaluation. *Health Econ* 1996;**5**:279–96.

405. Bleichrodt H, Johannesson M. An experimental test of a theoretical foundation for rating scale valuation. *Med Decis Making* 1997;**17**:208–16.

406. Torrance GW. Utility approach to measuring health-related quality of life [review]. *J Chron Dis* 1987;**40**:593–603.

407. Kaplan RM, Feeny D, Revicki DA. Methods for assessing relative importance in preference based outcome measures [review]. *Qual Life Res* 1993;**2**:467–75.

408. Busschbach JJ, Horikx PE, van den Bosch JM, Brutel dlR, de Charro FT. Measuring the quality of life before and after bilateral lung transplantation in patients with cystic fibrosis. *Chest* 1994;**105**:911–17.

409. Bakker C, Rutten M, van Doorslaer E, Bennett K, van der Linden S. Feasibility of utility assessment by rating scale and standard gamble in patients with ankylosing spondylitis or fibromyalgia. *J Rheumatol* 1994;**21**:269–74.

410. Sivertssen E, Fjeld NB, Abdelnoor M. Quality-of-life after open-heart-surgery. *Vasc Surg* 1994;**28**:581–8.

411. Lenert LA, Morss S, Goldstein MK, Bergen MR, Faustman WO, Garber AM. Measurement of the validity of utility elicitations performed by computerized interview. *Med Care* 1997;**35**:915–20.

412. Revicki DA. Relationship between health utility and psychometric health status measures. *Med Care* 1992;**30** Suppl 5:MS274–82.

413. Morss SE, Lenert LA, Faustman WO. The side effects of antipsychotic drugs and patients' quality of life: patient education and preference assessment with computers and multimedia. Proceedings of the Annual Symposium on Computer Applications in Medical Care; 1993. p. 17–21.

414. Van der Donk J, Levendag PC, Kuijpers AJ, Roest FHJ, Habbema JDF, Meeuwis CA, *et al.* Patient participation in clinical decision-making for treatment of T3 laryngeal cancer: a comparison of state and process utilities. *J Clin Oncol* 1995;**13**:2369–78.

415. Wolfson AD, Sinclair AJ, Bombardier C, McGreer A. Preference measurements for functional status in stroke patients: interrater and intertechnique comparisons. In: Kane RL, Kane RA, editors. Values and long term care. Lexington, MA: Lexington Books; 1982. p. 191–214.

416. Shiell A, King M, Briggs A. The consistency of rating scale and time trade off techniques for eliciting preference weights for health states. Paper presented at the UK Health Economists' Study Group Meeting; 1993 Jul; Glasgow.

417. Krabbe PFM, EssinkBot ML, Bonsel GJ. The comparability and reliability of five health-state valuation methods. *Soc Sci Med* 1997;**45**:1641–52.

418. O'Connor AM. Validation of a decisional conflict scale. *Med Decis Making* 1995;**15**:25–30.

419. O'Brien B, Viramontes JL. Willingness-to-pay – a valid and reliable measure of health state preference. *Med Decis Making* 1994;**14**:289–97.

420. Gabriel SE, Campion ME, O'Fallon WM. Patient preferences for nonsteroidal antiinflammatory drug related gastrointestinal complications and their prophylaxis. *J Rheumatol* 1993;**20**:358–61.

421. Sutherland HJ, Lockwood GA, Minkin S, Tritchler DL, Till JE, Llewellyn-Thomas HA. Measuring satisfaction with health care: a comparison of single with paired rating strategies. *Soc Sci Med* 1989;**28**:53–8.

422. Rutten-van Molken MP, Bakker CH, van Doorslaer EK, van der Linden S. Methodological issues of patient utility measurement. Experience from two clinical trials. *Med Care* 1995;**33**:922–37.

423. Pinto PJ. Is the person trade-off a valid method for allocating health care resources? *Health Econ* 1997;**6**:71–81.

424. O'Leary JF, Fairclough DL, Jankowski MK, Weeks JC. Comparison of time-tradeoff utilities and rating scale values of cancer patients and their relatives: evidence for a possible plateau relationship. *Med Decis Making* 1995;**15**:132–7.

425. Parducci A. Contextual effects. A range–frequency analysis. In: Cartarette E, Friedman M, editors. Handbook of perception. Vol II. New York: Academic Press; 1974. p. 127–41.

426. Dyer JS, Sarin RK. Relative risk-aversion. *Manag Sci* 1982;**28**:875–86.

427. Anderson N. Functional measurement and psychophysical judgement. *Psychol Rev* 1970;**77**:153–70.

428. Anderson N. Integration theory and attitude change. *Psychol Rev* 1971;**78**:171–206.

429. Appels A, Bosma H, Grabauskas V, Gostautas A, Sturmans F. Self-rated health and mortality in a Lithuanian and a Dutch population. *Soc Sci Med* 1996;**42**:681–9.

430. Adelman RD, Fields SD, Jutagir R. Geriatric education. 2. The effect of a well elderly program on medical-student attitudes toward geriatric-patients. *J Am Geriatr Soc* 1992;**40**:970–3.

431. Avis M. Incorporating patients' voices in the audit process. *Qual Health Care* 1997;**6**:86–91.

432. Carr-Hill R. The measurement of patient satisfaction. *J Public Health Med* 1992;**14**:236–49.

433. Cleary PD. The importance of patient surveys. *BMJ* 1999;**319**:720–1.

434. Locker D, Dunt D. Theoretical and methodological issues in sociological studies of consumer satisfaction with medical care. *Soc Sci Med* 1978;**12**:283–92.

435. Williams B. Patient satisfaction: a valid concept? *Soc Sci Med* 1994;**38**:509–16.

436. McKinley RK, Manku-Scott T, Hastings AM, French DP, Baker R. Reliability and validity of a new measure of patient satisfaction with out of hours primary medical care in the UK: development of a patient questionnaire. *BMJ* 1997;**314**:193–8.

437. Cohen G, Forbes J, Garraway M. Can different patient satisfaction survey methods yield consistent results? Comparison of three surveys. *BMJ* 1996;**313**:841–4.

438. Avis M. Incorporating patients' voices in the audit process. *Qual Health Care* 1997;**6**:86–91.

439. Porter M, MacIntyre S. What is, must be best: a research note on conservative or differential responses to antenatal care provision. *Soc Sci Med* 1984;**19**:1197–200.

440. Fitzpatrick R, Hopkins A. Problems in the conceptual framework of patient satisfaction research. *Sociol Health Illn* 1983;**5**:297–311.

441. Cleary PD, McNeill BJ. Patient satisfaction as an indicator of quality care. *Inquiry* 1998;**25**:25–36.

442. Oswald SL, Turner DE, Snipes RL, Butler D. Quality determinants and hospital satisfaction. Perceptions of the facility and staff might be key influencing factors. *Mark Health Serv* 1998;**18**:18–22.

443. Mishra DP, Singh J, Wood V. An empirical investigation of two competing models of patient satisfaction. *J Ambul Care Mark* 1991;**4**(2):17–36.

444. Shewchuk RM, O'Connor SJ, White JB. In search of service quality measures: some questions regarding psychometric properties. *Health Serv Manage Res* 1991;**4**:65–75.

445. Tomes A, Ng S. Service quality in hospital care: the development of an in-patient questionnaire. *Int J Health Care Qual Assur* 1995;**8**(3):25–33.

446. Youssef F, Nel D, Bovaird T. Service quality in NHS hospitals. *J Manag Med* 1995;**9**:66–74.

447. Hart MC. Measuring perception of quality in NHS clinics using SERVQUAL methodology. In: Richards B, editor. Current perspectives in healthcare computing. Weybridge, Surrey: BJHC; 1996. p. 37–42.

448. Dyck D. Gap analysis of health services. Client satisfaction surveys [review]. *AAOHN J* 1996;**44**(11):541–9.

449. Ryan M. Using consumer preferences in health care decision-making: the application of conjoint analysis. London: Office of Health Economics; 1997.

450. McFadden D. Conditional logit analysis of qualitative choice behavior. Berkeley, CA: University of California at Berkeley; 1973.

451. Hannemann W. Welfare evaluations in contingent valuation experiments with discrete choices: reply. *Am J Agric Econ* 1984;**69**:332–41.

452. Peralta-Carcelen MFC, Coston D, Dolan JG. Preferences of pregnant women and physicians for two strategies for prevention of early onset group B streptococcal sepsis in neonates. *Arch Pediatr Adolesc Med* 1997;**151**:712–18.

453. Dolan JG, Bordley DR. Individualized patient decision making using the analytic hierarchy process (AHP): reliability, validity, and clinical usefulness [meeting abstract]. *Med Decis Making* 1991;**11**:322.

454. Dolan JG, Bordley DR. Bedside decision making using the analytic hierarchy process. *Med Decis Making* 1992;**12**:344.

455. Dolan JG. The role of diagnostic endoscopy in the management of low risk patients with acute pyelonephritis [meeting abstract]. *Clin Res* 1992;**40**:580A.

456. Blumenschein K, Johannesson M. Relationships between quality of life instruments, health state utilities, and willingness to pay in patients with asthma. *Annal Allergy Asthma Immunol* 1998;**80**:189–94.

457. Torrance GW, Feeny DH, Furlong WJ, Barr RD, Zhang Y, Wang Q. Multiattribute utility function for a comprehensive health status classification system – Health Utilities Index Mark 2. *Med Care* 1996;**32**:702–22.

458. Shackley P, Cairns J. Evaluating the benefits of antenatal screening: an alternative approach. *Health Policy* 1996;**36**:103–15.

459. Gage BF, Cardinalli AB, Owens DK. The effect of stroke and stroke prophylaxis with aspirin or warfarin on quality of life. *Arch Intern Med* 1996;**156**:1829–36.

460. Rabin R, Rosser RM, Butler C. Impact of diagnosis on utilities assigned to states of illness. *J R Soc Med* 1993;**86**:444–8.

461. Ramsey SD, Patrick DL, Lewis S, Albert RK, Raghu G. Improvement in quality of life after lung transplantation: a preliminary study. The University of Washington Medical Center Lung Transplant Study Group. *J Heart Lung Transplant* 1995;**14**:870–7.

462. Dolan P, Gudex C, Kind P, Williams A. Valuing health states: a comparison of methods. *J Health Econ* 1996;**15**:209–31.

463. Patrick DL, Starks HE, Cain KC, Uhlmann RF, Pearlman RA. Measuring preferences for health states worse than death. *Med Decis Making* 1994;**14**:9–18.

464. Hall J, Gerard K, Salkeld G, Richardson J. A cost utiliy analysis of mammography screening in Australia. *Soc Sci Med* 1992;**34**:993–1004.

465. Thompson MS. Willingness to pay and accept risks to cure chronic disease. *Am J Public Health* 1986;**76**:392–6.

466. Reed WW, Herbers JEJ, Noel GL. Cholesterol-lowering therapy: what patients expect in return. *J Gen Intern Med* 1993;**8**:591–96.

467. Torrance GW. Measurement of health state utilities for economic appraisal [review]. *J Health Econ* 1986;**5**:1–30.

468. Bosch JL, Hunink MG. The relationship between descriptive and valuational quality-of-life measures in patients with intermittent claudication [review]. *Med Decis Making* 1996;**16**:217–25.

469. Zug KA, Littenberg B, Baughman RD, Kneeland T, Nease RF, Sumner W, *et al.* Assessing the preferences of patients with psoriasis. A quantitative, utility approach. *Arch Dermatol* 1995;**131**:561–8.

470. Bala MV, Wood LL, Zarkin GA, Norton EC, Gafni A, O'Brien B. Valuing outcomes in health care: a comparison of willingness to pay and quality-adjusted life years. *J Clin Epidemiol* 1998;**51**:667–76.

471. Scott A. Giving things up to have more of others. The implications of limited substitutability for eliciting preferences in health and health care. Aberdeen: Health Economics Research Unit; 1998. HERU Discussion Papers 01/98.

472. Von Winterfeldt D. Additivity and expected utility in risky multiattribute preferences. *J Math Psychol* 1980;**21**:66–82.

473. Tversky A, Sattath S, Slovic P. Contingent weighting in the judgement of choice. *Psychol Rev* 1988;**95**:371–84.

474. Slovic P, Tversky A. Who accepts Savage's axiom? *Behav Sci* 1974;**19**:368–73.

475. Kahneman D, Tversky A. Prospect theory: an analysis of decision making under risk. *Econometrica* 1979;**47**:263–92.

476. Bell DE. Regret in decision making under uncertainty. *Oper Res* 1982;**30**:961–81.

477. Loomes G, Sugden R. Regret theory: an alternative theory of rational choice under uncertainty. *Econ J* 1982;**92**:805–24.

478. Slovic P, Lichtenstein S. Preference reversals: a broader perspective. *Am Econ Rev* 1983;**73**:596–605.

479. Loomes G, Sugden R. The importance of what might have been. In: Hagen O, Wenstop F, editors. Progress in utility and risk theory. Dordrecht: Reidel; 1984. p. 219–35.

480. Bell DE. Disappointment in decision making under uncertainty. *Oper Res* 1985;**33**:1–27.

481. Loomes G, Sugden R. Disappointment and dynamic consistency in choice under uncertainty. *Rev Econ Stud* 1986;**53**:271–82.

482. Ryan M. Valuing psychological factors in the provision of assisted reproductive techniques using the economic instrument of willingness to pay. *J Econ Psychol* 1998;**19**:179–204.

483. Smith RD. Is regret theory an alternative basis for estimating the value of healthcare interventions? *Health Policy* 2000;**37**:105–15.

484. Fryback DG, Dasbach EJ, Klein R, Klein BE, Dorn N, Peterson K, *et al.* The Beaver Dam Health Outcomes Study: initial catalog of health-state quality factors. *Med Decis Making* 1993;**13**:89–102.

485. Johnson ES, Sullivan SD, Mozaffari E, Langley PC, Bodsworth NJ. A utility assessment of oral and intravenous ganciclovir for the maintenance treatment of AIDS-related cytomegalovirus retinitis. *Pharmacoeconomics* 1996;**10**:623–9.

486. Krumins PE, Fihn SD, Kent DL. Symptom severity and patients' values in the decision to perform a transurethral resection of the prostate. *Med Decis Making* 1988;**8**:1–8.

487. Detsky AS, McLaughlin JR, Abrams HB, L'Abbe KA, Whitwell J, Bombardier C, *et al.* Quality of life of patients on long-term total parenteral nutrition at home. *J Gen Intern Med* 1986;**1**:26–33.

488. Glasziou PP, Bromwich S, Simes RJ. Quality of life six months after myocardial infarction treated with thrombolytic therapy. AUS-TASK Group. Australian arm of International tPA/SK Mortality Trial. *Med Jo Aust* 1994;**161**:532–6.

489. Kreibich DN, Vaz M, Bourne RB, Rorabeck CH, Kim P, Hardie R, *et al.* What is the best way of assessing outcome after total knee replacement? *Clin Orthop* 1996;**331**:221–5.

490. Handler RM, Hynes LM, Nease RFJ. Effect of locus of control and consideration of future consequences on time tradeoff utilities for current health. *Qual Life Res* 1997;**6**:54–60.

491. Churchill DN, Torrance GW, Taylor DW. Measurement of quality of life in end-stage renal disease: the time trade-off approach. *Clin Invest Med* 1987;**10**(1):14–20.

492. Molzahn AE, Northcott HC, Hayduk L. Quality of life of patients with end stage renal disease: a structural equation model. *Qual Life Res* 1996;**5**:426–32.

493. Kemp S. Magnitude estimation of the utility of public-goods. *J Appl Psychol* 1991;**76**:533–40.

494. Robinson A, Dolan P, Williams A. Valuing health status using VAS and TTO: what lies behind the numbers? *Soc Sci Med* 1997;**45**:1289–97.

495. Zethraeus N, Johannesson M, Henriksson P, Strand RT. The impact of hormone replacement therapy on quality of life and willingness to pay. *Br J Obstet Gynaecol* 1997;**104**:1191–5.

496. Swan JS, Fryback DG, Lawrence WF, Katz DA, Heisey DM, Hagenauer ME, *et al.* MR and conventional angiography: work in progress towards assessing utility in radiology. *Acad Radiol* 1999;**4**:475–82.

497. Sackett DL, Torrance GW. The utility of different health states as perceived by the general public. *J Chronic Dis* 1978;**31**:697–704.

498. Lipscomb J. Value preferences for health: meaning, measurement, and use in program evaluation. In: Kane RL, Kane RA, editors. Values and long term care. Lexington, MA: Lexington Books; 1982. p. 27–83.

499. Sutherland HJ, Llewellyn-Thomas H, Boyd NF, Till JE. Attitudes toward quality of survival. The concept of "maximal endurable time". *Med Decis Making* 1982;**2**:299–309.

500. Wakker P, Stigglebout A. Explaining distortions in utility elicitation through the rank dependent model for risky choices. *Med Decis Making* 1995;**15**:180–5.

501. Irvine EJ. Quality of Life in inflammatory bowel disease: biases and other factors affecting scores. *Scand J Gastroenterol Suppl* 1995;**208**:136–40.

502. Dolan P, Gudex C. Time preference, duration and health state valuations. *Health Econ* 1995;**4**:289–99.

503. Green C, Brazier J, Deverill M. A review of the evidence surrounding health state valuation techniques (VAS, TTO, SG, PTO, ME). Paper presented at the UK Health Economists' Study Group Meeting; 1998 Jul; Galway.

504. Nord E, Richardson J, Macarounas-Kirchmann K. Social evaluation of health care versus personal evaluation of health states. Evidence on the validity of four health-state scaling instruments using Norwegian and Australian surveys. *Int J Technol Assess Health Care* 1993;**9**:463–78.

505. Gafni A. Willingness to pay in the context of an economic evaluation of healthcare programs: theory and practice. *Am J Manag Care* 1997; **3** Suppl:S21–S32.

506. Pauly MV. Valuing health care benefits in money terms. In: Sloan FA, editor. Valuing health care: costs, benefits, and effectiveness of pharmaceuticals and other medical technologies. Cambridge: Cambridge University Press; 1995. p. 99–124.

507. Donaldson C, Farrar S, Mapp T, Walker A, Macphee S. Assessing community values in health care: is the 'willingness to pay' method feasible? *Health Care Anal* 1997;**5**:7–29.

508. Thompson MS, Read JL, Liang M. Feasibility of willingness-to-pay measurement in chronic arthritis. *Med Decis Making* 1984;**4**:195–215.

509. Kartman B, Stalhammar NO, Johannesson M. Contingent valuation with an open-ended follow-up question: a test of scope effects. *Health Econ* 1997;**6**:637–9.

510. Thompson MS, Read JL, Laing M. Willingness to pay concepts for social decisions in health. In: Kane RL, Kane RA, editors. Values and long term care. Lexington, MA. Lexington Books; 1982. p. 103–26.

511. Donaldson C, Hundley V, Mapp T. Willingness to pay: a method for measuring preferences for maternity care? *Birth* 1998;**25**:32–9.

512. Donaldson C, Thomas R, Torgerson DJ. Validity of open-ended and payment scale approaches to eliciting willingness to pay. *Appl Econ* 1997;**29**:79–84.

513. Johannesson M, Johansson P, Kristrom B, Borgquist L, Jönsson B. Willingness to pay for lipid lowering: a health production function approach. *Appl Econ* 1993;**25**:1023–31.

514. Eckerlund I, Johannesson M, Johansson P-O, Tambour M, Zethraeus N. Value for money? A contingent valuation study of the optimal size of the Swedish health care budget. *Health Policy* 1999;**34**:135–43.

515. Johannesson M, O'Conor RM, Kobelt-Nguyen G, Mattiasson A. Willingness to pay for reduced incontinence symptoms. *Br J Urol* 1997;**80**:557–62.

516. Ryan M. Using willingness to pay to assess the benefits of assisted reproductive techniques. *Health Econ* 1996;**5**:543–58.

517. Ryan M. Should government fund assisted reproductive techniques? A study using willingness to pay. *Appl Econ* 1997;**29**:841–9.

518. Lindholm L, Rosén ME, Stenbeck ME. Determinants of willingness to pay taxes for a community-based prevention programme. *Scand J Soc Med* 1997;**25**:126–35.

519. Flowers CR, Garber AM, Bergen MR, Lenert LA. Willingness-to-pay utility assessment: feasibility of use in normative patient decision support systems. *J Am Med Informatic Assoc Proc* 1997;223–7.

520. Coley CM, Li Y-H, Medsger AR, Marrie TJ, Fine MJ, Kapoor WN, *et al.* Preferences for home vs. hospital care among low-risk patients with community-acquired pneumonia. *Arch Intern Med* 1996;**156**:1565–71.

521. Dranitsaris G. A pilot study to evaluate the feasibility of using willingness to pay as a measure of value in cancer supportive care: an assessment of amistofine cytoprotection. *Support Care Cancer* 1997;**5**:489–99.

522. Lee SJ, Neumann PJ, Churchill WH, Cannon ME, Weinstein MC, Johannesson M. Patients' willingness to pay for autologous blood donation. *Health Policy* 1997;**40**:1–12.

523. Johannesson M, Åberg H, Agréus L, Borgquist L, Jönsson B. Cost–benefit analysis of non-pharmacological treatment of hypertension. *J Intern Med* 1991;**230**:307–12.

524. Johannesson M. Economic evaluation of lipid lowering – a feasibility test of the contingent valuation approach. *Health Policy* 1992;**20**:309–20.

525. Van der Pol M, Cairns J. Eliciting individual time preferences for own health using a dichotomous choice question with follow-up. Paper presented to the 4th Nordic Health Econometrics Workshop; 1998 Aug 20; Oslo.

526. Schkade DA, Payne JW. How people respond to contingent valuation questions: a verbal protocol analysis of willingness to pay for an environmental regulation. *J Environ Econ Manage* 1994;**26**:88–109.

527. Loehman ET, Berg SV, Arroyo AA, Hedinger RA, Schwartz JM, Shaw ME, *et al.* Distributional analysis of regional benefits and cost of air quality control. *J Environ Econ Manage* 1979;**6**:222–43.

528. Pennie RA, O'Connor A, Garvock M, Drake E. Factors influencing the acceptance of Hepatitis B vaccine by students in health disciplines in Ottawa. *Can J Public Health* 1991;**82**:12–15.

529. Asenso-Okyere WK, Osei-Akoto I, Anum A, Appiah EN. Willingness to pay for health insurance in a developing country – a pilot study of the informal sector of Ghana using contingent valuation. *Health Policy* 1997;**42**:223–37.

530. Appel L, Steinberg E, Powe N, Anderson G, Dwyer S, Faden R. Risk reduction from low osmality contrast media: what do patients think it's worth? *Med Care* 1990;**28**:324–37.

531. Berwick DM, Weinstein MC. What do patients value? Willingness to pay for ultrasound in normal pregnancy. *Med Care* 1985;**23**:881–93.

532. Anderson G, Black C, Dunn E, Alonso J, Christian-Norrega J, Folmer-Anderson T, *et al.* Willingness to pay to shorten waiting time for cataract surgery. *Health Aff* 1997;**16**:181–91.

533. Acton J. Evaluating public programmes to save lives: the case of heart attacks. Santa Monica, CA: RAND Corporation; 1973. Report no. R950RC.

534. O'Brien B, Goeree R, Gafni A, Torrance GW, Pauly MV, Erder H, *et al.* Assessing the value of a new pharmaceutical: a feasibility study of contingent valuation in managed care. *Med Care* 1998;**36**:370–84.

535. Ramsey SD, Sullivan SD, Psaty BM, Patrick DL. Willingness to pay for antihyperintensive care: evidence from a staff-model HMO. *Soc Sci Med* 1997;**44**:1911–17.

536. Gore P, Madhavan S. Consumers' preference and willingness to pay for pharmacist counselling for non-prescription medicines. *J Clin Pharm Ther* 1994;**19**:12–25.

537. Johannesson M, Johansson B. On the value of changes in life-expectancy: blips versus parametric changes. *J Risk Uncertainty* 1997;**15**:221–39.

538. O'Brien B, Novosel S, Torrance G, Streiner D. Assessing the economic value of a new antidepressant: a willingness to pay approach. *Pharmacoeconomics* 1995;**8**:34–45.

539. Osmond M, Klassen T, Quinn J. Economic comparison of a tissue adhesive and suturing in the repair of pediatric facial lacerations. *J Pediatr* 1995;**126**:892–5.

540. Reutzel TJ, Furmaga E. Willingness to pay for pharmacist services in a veterans administration hospital. *J Res Pharm Econ* 1993;**5**(2):89–114.

541. Burgoyne CB. Distributive justice and rationing in the NHS: framing effects in press coverage of a controversial decision. *J Community Appl Soc Psychol* 1997;**7**:119–36.

542. Fox-Rushby J. Willingness to pay as a method for valuing health related quality of life. Paper presented at the UK Health Economists' Study Group Meeting; 1991 Jul; Aberdeen.

543. Granberg M, Wikland M, Nilsson L, Hamberger L. Couples' willingness to pay for IVF/ET. *Acta Obstet Gynecol Scand* 1995;**74**:199–202.

544. Walraven G. Willingness to pay for district hospitals in rural Tanzania. *Health Policy Plan* 1996;**11**:428–37.

545. Fischer GW. Willingness to pay for probibalistic improvements in health status: a psychological perspective. In: Mushkin SJ, Dunlop DW, editors. Health: what is it worth? New York: Pergamon; 1979. p. 167–200.

546. Garbacz C, Thayer M. An experiment in valuing senior companion program services. *J Hum Resour* 1983;**18**:147–53.

547. Kobelt G. Economic considerations and outcome measurement in urge incontinence. *Urology* 1997;**50**(6A):100–7.

548. Muller A, Reutzel TJ. Willingness to pay for reduction in fatality risk: an exploratory survey. *Am J Public Health* 1984;**74**:808–12.

549. Baron J. Confusion of relative and absolute risk in valuation. *J Risk Uncertainty* 1997;**14**:301–9.

550. Gyldmark M. Preferences for health care services in Denmark. An investigation through contingent valuation. Paper presented at the 14th Meeting of the Nordic Health Economists Group; 1993 Aug 18–20; Tromso.

551. Olsen JA. Aiding priority setting in health care: is there a role for the contingent valuation method? [review]. *Health Econ* 1997;**6**:603–12.

552. Kartman B, Andersson F, Johannesson M. Willingness to pay for reductions in angina pectoris attacks. *Med Decis Making* 1996;**16**:248–53.

553. Kartman B, Stalhammar NO, Johannesson M. Valuation of health changes with the contingent valuation method: a test of scope and question order effects. *Health Econ* 1996;**5**:531–41.

554. Viscusi WK, Magat WA, Huber J. Pricing environmental health risks: survey assessments of risk–risk and risk–dollar trade-offs for chronic bronchitis. *J Environ Econ Manage* 1991;**21**:32–51.

555. Ryan M, Scott DA, Donaldson C. Valuing health care using willingness to pay: a comparison of the payment card and dichotomous choice methods. Paper presented to the 2nd International Conference of the International Health Economics Association; 1999 Jun 7–9; Rotterdam.

556. Johannesson M. Economic evaluation of hypertension treatment. *Int J Technol Assess Health Care* 1992;**8**:506–23.

557. Lee SJ, Liljas B, Neumann PJ, Weinstein MC, Johannesson M. The impact of risk information on patients' willingness to pay for autologous blood donation. *Med Care* 1998;**36**:1162–73.

558. Stalhammar NO. An empirical note on willingness to pay and starting-point bias. *Med Decis Making* 1996;**16**:242–7.

559. Ryan M, Scott DA, Donaldson C. Econometric issues raised in the analysis of payment scale and closed-ended contingent valuation data sets in health care. Paper presented to the 5th Nordic Health Econometrics Workshop; 1998 Aug 20; Oslo.

560. Ryan M, Ratcliffe J. Some issues in the application of closed-ended willingness to pay studies to valuing health goods: an application to antenatal care in Scotland. *Appl Econ* 2000;**32**:643–51.

561. Mullen PM. Public involvement in health-care priority setting: an overview of methods for eliciting values. *Health Expectations* 1999;**2**:222–34.

562. Srivastava J, Connolly T, Beach LR. Do ranks suffice? A comparison of alternative weighting approaches in value elicitation. *Organ Behav Hum Decis Process* 1995;**63**:112–16.

563. Mullen P, Spurgeon P. Priority setting and the public. Abingdon, Oxon: Radcliffe Medical Press; 2000.

564. Hoinville G, Courtenay G. Measuring consumer priorities. In: O'Riordan T, D'Arge RC, editors. Progress in resource management and environmental planning. Vol. 1. Chichester: Wiley; 1979. p. 143–70.

565. Browne JP, McGee HM, O'Boyle CA. Conceptual approaches to the assessment of quality of life. *Psychol Health* 1997;**12**:737–51.

566. Bowling A, Jacobson B, Southgate L. Health-service priorities – explorations in consultation of the public and health-professionals on priority setting in an inner London health district. *Soc Sci Med* 1993;**37**:851–7.

567. Clarke R. Developments in deliberative approaches. Paper presented at a seminar, 'New methods for involving the public in decisions'; 1999 Sep; Birmingham.

568. Holloway I. Basic concepts for qualitative research. Oxford: Blackwell Science; 1997.

569. The Penguin dictionary of sociology. 2nd ed. Harmondsworth: Penguin Books; 1988.

570. Murphy MK, Black NA, Lamping DL, McKee CM, Sanderson CFB, Askham J, *et al.* Concensus development methods, and their use in clinical guideline development. *Health Technol Assess* 1998;**2**(3).

571. Williams PL, Webb C. The Delphi technique: a methodological discussion. *J Adv Nurs* 1994;**19**:180–6.

572. Donovan J, Coast J. Public preferences in priority setting – unresolved issues. In: Malek M, editor. Setting priorities in health care. Chichester: John Wiley & Sons, 1994; p. 31–45.

573. Hemenway D, Killen A. Complainers and noncomplainers. *J Ambulatory Care Manage* 1989;**12**(3):19–27.

574. Kitzhaber JA. Prioritising health services in an era of limits: the Oregon experience. *BMJ* 1993;**307**:377.

575. Dean H, Gale K, Woods R. 'This isn't very typical I'm afraid': observing community care complaints procedures. *Health Soc Care Community* 1996;**4**:338–46.

576. Eriksson CG. Focus groups and other methods for increased effectiveness of community intervention – a review [review]. *Scand J Primary Health Care Suppl* 1988;**1**:73–80.

577. Beaudin CL, Pelletier LR. Consumer-based research: using focus groups as a method for evaluating quality of care. *J Nurs Care Qual* 1996;**10**(3):28–33.

578. Tang KC, Davis A. Critical factors in the determination of focus group size [review]. *Fam Pract* 1995;**12**:474–5.

579. Birdwell SW, Calesaric H. Identifying health care needs of rural Ohio citizens: an evaluation of a two-stage methodology. *J Rural Health* 1996;**12**:130–6.

580. Merton GB, Levine RJ, Koocher GP, Rosenthal R, Thompson WC. Community consultation in socially sensitive research: lessons from clinical trials for treatment of AIDS. *Am Psychol* 1988;**43**:573–81.

581. Price D. Choices without reasons: citizens' juries and policy evaluation. *J Med Ethics* 2000;**26**:272–6.

582. Seiler H. Review of "Planning Cells:" problems of legitimation. In: Renn O, Webler T, Wiedemann P, editors. Fairness and competition in citizen participation: evaluating models for environmental discourse. Dordrecht: Kluwer Academic; 1995. p. 144–51.

583. Lupton C, Peckham S, Taylor P. Managing public involvement in healthcare purchasing. Buckingham: Open University Press; 1998.

584. Chess C, Purcell K. Public participation and the environment: do we know what works? *Environ Sci Technol* 1999;**33**:2685–92.

585. Jessop EG. Public meetings deserve proper science. *J Public Health Med* 1999;**21**:365–6.

586. Kocur G, Alder T, Hyman W, Aunet B. Guide to forecasting travel demand with direct utility assessment: Final report to the US Department of Transportation. Hanover, NH: Resource Policy Center, Thayer School of Engineering, Dartmouth College; 1982.

587. Cookson R. Incorporating psychosocial considerations into health valuation: an experimental study. *J Health Econ* 2000;**19**:369–401.

588. Majid I, Sinden JA, Randall A. Benefit valuation of increments to existing systems of public facilities. *Land Econ* 1983;**59**:377–92.

589. Hoehn J. Valuing the multidimensional impacts of environmental policy: theory and methods. *Am Agric Assoc* 1991;**73**:289–99.

590. Hoehn J, Randall A. Too many proposals pass the benefit–cost test. *Am Econ Rev* 1989;**79**:544–51.

591. Gafni A. Willingness-to-pay as a measure of benefits: relevant questions in the context of public decision making about health care programmes. *Med Care* 1991;**29**:1246–52.

592. Benbassat J, Pilpel D, Tidhar M. Patients' preferences for participation in clinical decision making: a review of published studies. *Behav Med* 1998;**24**:81–8.

593. Guadagnoli E, Ward P. Patient participation in decision-making. *Soc Sci Med* 1998;**47**:329–39.

594. Heginbotham C. Health care priority setting: a survey of dcotors, managers, and the general public. In: Smith R, editor. Rationing in action. London: BMJ Publishing; 1993. p. 141–56.

595. Bowling A. Health care rationing: the public's debate. *BMJ* 1996;**312**:670–4.

596. Kneeshaw J. What does the public think about rationing? A review of the evidence. In: New B, editor. Rationing: talk and action in health care. London: King's Fund/BMJ Publishing Group; 1997. p. 58–76.

597. Coast J. Rationing within the NHS should be explicit: the case against. *BMJ* 2000;**314**:1118–22.

598. Coast J. Explicit rationing, deprivation disutility and denial disutility: evidence from a qualitative study. In: Coulter A, Ham C, editors. The global challenge of health care rationing. Buckingham: Open University Press; 2000. p. 192–200.

599. Fischhoff B. Value elicitation: is there anything in there? *Am Psychol* 1991;**46**:835–47.

600. Gregory R, Lichestein S, Slovic P. Valuing environmental resources: a constructive approach. *J Risk Uncertainty* 1993;**7**:177–97.

601. Slovic P. The construction of preference. *Am Psychol* 1995;**50**:364–71.

602. Shiell A, Hawe P, Seymour J. Values and preferences are not necessarily the same. *Health Econ* 1997;**6**:515–18.

603. Snowdon H, Jones G. Local advisory groups in the Sedgefield area. Paper presented at a seminar, 'New methods for involving the public in decisions'; 1999 Sep; Birmingham.

604. Cartwright A. The dignity of labour: a study of childbearing and induction. London: Tavistock; 1979.

605. Bate A, Ryan M. Examining patient preferences for junior doctors versus specialist nurses in the provision of rheumatology services: a discrete choice experiment. Aberdeen: University of Aberdeen; 1998. HERU Discussion Paper 05/98.

606. Salkeld G, Ryan M, Short L. The veil of experience: do consumers prefer what they know best? *Health Econ Lett* 2000;**4**:4–9.

607. Thayler R. Toward a positive theory of consumer choice. *J Behav Organ* 1980;**1**:39–60.

608. Samuelson W, Zeckhauser R. Status quo bias in decision making. *J Risk Uncertainty* 1988;**1**:7–59.

609. Knetsch JL, Sinden JA. Wilingness to pay and compensation demanded: experimental evidence of an unexpected disparity in measures of value. *Q J Econ* 1984;**99**:507–21.

610. Knetsch JL, Sinden JA. The persistence of evaluation disparities. *Q J Econ* 1987;**102**:691–5.

611. Knetsch JL. The endowment effect and evidence of non-reversible indifference curves. *Am Econ Rev* 1989;**790**:1277–84.

612. Morrison GC. Willingness to pay and willingness to accept: some evidence of an endowment effect. *Appl Econ* 1997;**29**:411–17.

613. Morrison GC. Resolving differences in willingness to pay and willingness to accept: comment. *Am Econ Rev* 1997;**87**:236–40.

614. Klein L. "Planning for Real" and other methods of public participation. Paper presented at a seminar, 'New methods for involving the public in decisions'; 1999 Sep; Birmingham.

615. Tashhakkori A, Teddlie C. Mixed methodology: combining qualitative and quantitative approaches. London: Sage; 1998.

616. Ferraz MB, Quaresma MR, Goldsmith CH, Bennett K, Atra E. Corticosteroids in patients with rheumatoid-arthritis – utility. *Rev Rhum* 1994;**61**:255–9.

617. Read JL, Quinn RJ, Berwick DM, Fineberg HV, Weinstein MC. Preferences for health outcomes. Comparison of assessment methods. *Med Decis Making* 1984;**4**:315–29.

618. Clarke AE, Goldstein MK, Michelson D, Garber AM, Lenert LA. The effect of assessment method and respondent population on utilities elicited for Gaucher disease. *Qual Life Res* 1997;**6**:169–84.

619. Freeman JK, Szeinbach SL, Barnes JH, Garner DD, Gilbert FW. Assessing the need for student health services using maximum difference conjoint analysis. *J Res Pharm Econ* 1998;**9**(3):35–49.

620. Szeinbach SL, Barnes JH, McGhan WF, Murawski MM, Corey R. Using conjoint analysis to evaluate health state preferences. *Drug Inf J* 1999;**33**:849–58.

621. McNeil BJ, Pauker SG, Sox HCJ, Tvershy A. On the elicitation of preferences for alternative therapies. *N Engl J Med* 1982;**306**:1259–62.

622. Eckman MH. A counterpoint to the analytic hierarchy process. *Med Decis Making* 1989;**9**:57–8.

623. Loomes G, McKenzie L. The use of QALY's in health care decision making. *Soc Sci Med* 1989;**28**:299–308.

624. Laupacis A, Bourne R, Rorabeck C, Feeny D, Wong C, Tugwell P, *et al.* The effect of elective total hip replacement on health-related quality of life. *J Bone Joint Surg Am* 1993;**75**:1619–26.

625. Buckingham JK, Birdsall J, Douglas JG. Comparing three versions of the time tradeoff: time for a change? *Med Decis Making* 1996;**16**:335–47.

626. Mehrez A, Gafni A. Healthy-years equivalents versus quality-adjusted life years: in pursuit of progress. *Med Decis Making* 1993;**13**:287–92.

627. Redelmeier DA, Heller DN. Time preference in medical decision making and cost-effectiveness analysis. *Med Decis Making* 1993;**13**:212–17.

628. Nord E. The trade-off between severity of illness and treatment effect in cost–value analysis of health care. *Health Policy* 1993;**24**:227–38.

629. Donaldson C, Shackley P, Abdalla M, Miedzybrodzka Z. Willingness to pay for antenatal carrier screening for cystic fibrosis. *Health Econ* 1995;**4**:439–52.

630. Easthaugh S. Valuation of the benefits of risk-free blood. *Int J Technol Assess Health Care* 1991;**7**:51–7.

631. Mills A, Fox-Rushby J, Aikins M, D'Alessandro U, Cham K, Greenwood B. Financing mechanisms for village activities in The Gamba and their implications for financing insecticide for bednet impregnation. *J Trop Med Hyg* 1994;**97**:325–32.

632. Zeidner M, Shechter M. Reduction of test anxiety: a first attempt at economic evaluation. *Anxiety Stress Coping* 1994;**7**:1–18.

633. Bate A. Examining the importance of consumer views and preferences vis-a-vis other criteria that are commonly used to aid priority setting decision making: a pilot study [MSc Health Economics thesis]. York: University of York; 1999.

634. Silva MC, Lewis CK. Ethics, policy, and allocation of scarce resources in nursing service administration: a pilot study. *Nurs Connections* 1991;**4**(2):44–52.

635. Campbell SM, Hann M, Roland MO, Quayle JA, Shekelle PG. The effect of panel membership and feedback on ratings in a two-round Delphi survey: results of a randomized controlled trial. *Med Care* 1999;**37**:964–8.

636. Bradley M. Users manual for SPEED, version 2.1. The Hague: Hague Consulting Group; 1991.

637. Ham C. Priority setting in the NHS: reports from six districts. *BMJ* 1993;**307**:435–8.

638. Argyll and Clyde Health Board Priority Working Group. Prioritisation scoring index. Paisley: Department of Public Health, Argyll and Clyde Health Board; 1998.

639. Aryshire and Arran Health Board Spa Group. Setting Priorities in Ayrshire: report of the Ayrshire and Arran Health Board SPA Group. Ayr: Aryshire and Arran Health Board; 1999.

640. Cumming J. An overview of the health economics and health policy literatures on priority setting in health care. Wellington, New Zealand: Health Services Research Centre, University of Wellington; 1996. Health Services Research Centre Research Reports no. 4.

641. Dixon J, New B. Setting priorities New Zealand-style [editorial]. *BMJ* 1997;**314**:86–7.

642. New B, Le Grand J. Rationing in the NHS: principles and pragmatism. London: King's Fund; 1996.

643. Cumming J. Core services and priority-setting the New Zealand Experience. *Health Policy* 1994;**29**:41–60.

644. Dumfries and Galloway Health Board. Prioritisation ranking for service development proposals. Dumfries: Dumfries and Galloway Health Board; 1998.

645. Feighan T. Setting health care priorities in Northern Ireland. Work in progress paper presented at the UK Health Economists' Study Group Meeting; 1998 Jul; Galway.

646. Ham C. Priority setting in health care: learning from international experience. *Health Policy* 1997;**42**:49–66.

647. Honigsbaum F. Who shall live? Who shall die? – Oregon's health financing proposals. London: King's Fund College; 1993. King's Fund College Papers No. 4.

648. Hunter D. Desperately seeking solutions: rationing in health care. London: Longman; 1997.

649. Greenhalgh T. How to read a paper: the basics of evidence based medicine. London: BMJ Publishing Group; 1997.

650. Stevenson R, Hegarty M. In the picture. *Health Serv J* 1994;**104**:22–4.

651. Scottish Association of Health Councils. An analysis of health council effectiveness: representing the public in the NHS in Scotland. Edinburgh: Scottish Association of Health Councils; 1995.

652. Nitzan S. The vulnerability of point-voting schemes to preference variation and strategic manipulation. *Public Choice* 1985;**47**:349–70.

653. LeBreton M, Truchon M. A Borda measure for social choice functions. *Math Soc Sci* 1997;**34**:249–72.

Appendix 1

Electronic database search strategies for identifying methods for eliciting public preferences

MEDLINE and HealthSTAR (Ovid)

001 exp consumer participation/
002 exp consumer satisfaction/
003 public opinion/
004 ((public or consumer$ or patient$) adj3 (preference$ or opinion or choice$ or participat$)).tw
005 or/1–4
006 exp data collection/
007 exp research
008 elicit$.tw
009 measure$.tw
010 obtain$.tw
011 technique$.tw
012 or/6–11
013 5 and 12
014 exp health planning/
015 5 and 14
016 13 or 15

EMBASE (BIDS – Ovid)

001 ((public or consumer$ or patient$) adj3 (preference$ or opinion or choice$ or participat$)).tw
002 patient satisfaction/
003 patient attitude/
004 public opinion/
005 or/1–4
006 measurement/
007 exp information processing/
008 elicit$.tw
009 obtain$.tw
010 measure$.tw
011 technique$.tw
012 or/6–11
013 5 and 12
014 consumer/
015 healthcare planning/
016 resource allocation/
017 14 and (15 or 16)
018 13 or 17

Social Science Citation Index (BIDS)

001 (public or consumer* or patient*) @TKA
002 (preference* or choice or opinion or consult* or participat*) @TKA
003 (measur* or elicit* or obtain* or technique*) @TKA
004 1 and 2 and 3

PsycLIT (SilverPlatter)

001 explode "CONSUMER-RESEARCH"
002 "PUBLIC-OPINION"
003 "PREFERENCES"
004 (public or consumer? or patient?) near3 (preference? or choice? or opinion or consult* or participat*)
005 #1 or #2 or #3 or #4
006 explode "METHODOLOGY"
007 explode "MEASUREMENT"
008 (measure* or elicit* or obtain* or technique?)
009 #6 or #7 or #8
010 #5 and #9
011 explode "PREFERENCE-MEASURES"
012 public or consumer? or patient?
013 #11 and #12
014 #10 or #13

EconLIT (Ovid)

001 consumer choice.hw
002 ((public or consumer$ or patient$) adj3 (preference$ or opinion or choice$ or participat$)).tw
003 1 or 2
004 (elicit$ or measure$ or method$ or obtain$ or technique$).tw
005 3 and 4

Appendix 2

Additional data for chapter 5

Appendix 2 presents a summary of some of the studies identified in the review to provide the reader with a summary of the main issues raised for each of the techniques. This table does not provide a listing of all the studies identified in this review.

The only quantitative technique not mentioned in this summary is satisfaction surveys. There are two reasons for this. First, satisfaction has been ascertained using Likert scales and Guttman scales and these are covered in the tables. Secondly, our review of satisfaction surveys was at a general level, and readers are advised to refer to the body of the text.

Simple ranking exercises

Properties: there is no well-defined theoretical basis for simple ranking exercises and the output of such studies is ordinal.

Ref.	Empirical/methodological study	Acceptability to respondents	Cost	Internal consistency	Reproducibility	Validity
16	Bowling *et al*., 1993 Public ranking of 12 healthcare services in City and Hackney. Consisted of community group discussions, and a random sample of the public and doctors. More detail of these surveys in Bowling[366]	For community group, 9 of 359 did not complete questionnaire. For random sample of public, postal questionnaire initially gave response rate of 11%. Follow-up after 4 reminders, by interview, increased response rate to 78%. Response rates, after 4 mailings, was 66% (131/197) for the doctors, 68% (82/121) for the consultants and 6/7 of the public health doctors	Concern raised about high cost of public forums and interview surveys. Cited to range between £20,000 and £90,000			Given complexity of issues, personal interviews are the only valid approach
358	Shickle, 1997 Review of public preferences for prioritisation in the UK			Whilst respondents gave high rankings to life-saving treatment, in a binary choice they stated a preference for quality-of-life improving treatments		*A priori* hypotheses confirmed: preferences for life-saving treatments, treating the young, patients with dependants, and those who had led a healthy lifestyle
18	Furnham *et al*., 1998 Public prioritisation of kidney dialysis patients differing by sex, age, income, drinking habits and religious beliefs	Response rate by questionnaire taken home with patients was 81%				*A priori* hypotheses were confirmed – females were favoured over males, non-smokers over smokers, and poor people over rich. Findings agreed with Furnham and Briggs[26]
24	Angermeyer *et al*., 1999 Public attitudes towards treatment of depression. From 8 treatments, respondents were asked to prioritise top 2	Representative survey, response rate was 71.2% (1564 interviews)				Ranking led to more prominent differences between the potential sources of help than rating scales Strange that mental health professionals do not play a role in help-seeking in major depression Ranking is a closer reflection of reality since people are forced to make decisions as they do in daily life

continued

Simple ranking exercises *contd*

Properties: there is no well-defined theoretical basis for simple ranking exercises and the output of such studies is ordinal.

Ref.	Empirical/methodological study	Acceptability to respondents	Cost	Internal consistency	Reproducibility	Validity
25	Furnham, 1996 People's rank ordering (in terms of 4 dimensions: gender, income, alcohol consumption and religious beliefs) of treatment for patients suffering from kidney failure	95% (159/167) response rate				*A priori* hypotheses confirmed (and in line with other studies): respondents favoured females over males, poor over rich, and non-drinkers over drinkers.[18,26] They also favoured Christians over atheists
366	Bowling, 1999 More detailed report of City and Hackney public priority ranking of 12 healthcare services.[16] Community group discussions, a random sample of the public and doctors	Pilot indicated respondents found it difficult to rank all 12 options – therefore categorised into a 4-category ranking method ('essential', 'very important', 'important', 'less important') Main study comprised 350 respondents. Whilst response rate to the postal survey was poor,16 the follow-up by letter and interview was good. Response rates for the doctors were 66% for the consultants, 68% for the GPs and 6/7 of the public health doctors after 4 mailings 15 respondents gave negative or critical comments about the exercise. Respondents found the exercise very difficult and the wording may require further simplification Researchers may have to go out and interview people in deprived area rather than relying on them to return questionnaires Issue too complex for postal questionnaires	Study did not obtain the cooperation of the local health council who protested about the ethics of it The study (community groups, postal and interview surveys) took 12 months to complete and cost £24,000	The consistency of the public's prioritisation of mental health services as a medium priority was confirmed in a consistency check against additional attitude questions Doctors were asked an OE question to list in order of priority 5 areas of improvement they would like in City and Hackney. Their responses were consistent with their rankings	Not possible to compare pilot and main studies due to rewording in the questionnaire. Nevertheless, test–retest reliability was checked at the pilot stage and almost identical rankings were found	Higher priority for life-saving treatments and lower priority for health education and family planning consistent with earlier work[73] For community group, prioritisation of mental health and preventative services consistent with 5-item priority list (significant at 5% level) For random sample of public, mental health prioritisation, preventative services, and answers to the 5-item priorities list significantly correlated at 5% level Responses sensitive to question wording, e.g. in pilot "intensive care for premature babies" was given highest priority1 but when qualified with statement "weighing less than 11/2 pounds and unlikely to survive" rank dropped to 10

Qualitative discriminant process (QDP)

Properties: QDP has its theoretical basis in decision theory, fuzzy set theory, the theory of vague real numbers and voting theory. Given the mapping of qualitative responses onto an interval scale, the technique is held to measure strength of preference.

Ref.	Empirical/methodological study	Acceptability to respondents	Cost	Internal consistency	Reproducibility	Validity
28	Bryson *et al.*, 1994 Development of QDP method with illustration of identifying suitable candidates for the position of dean of the school of business	Authors note the technique is "simple and intuitively appealing". They also note its ease of use in software form	Computer-based so relatively expensive to collect data			
30	Ngwenyama & Bryson, 1998 Description of QDP method with illustrative example on the diagnosis, by 3 experts, of a patient whose symptoms indicate 4 possible diseases: gastric cancer, pancreatic cancer, functional disorder, and gallstones			One of the 3 experts gave inconsistent answers which, once brought to his/her attention, were corrected by the expert. Consistent answers are in this way assured with this method		

CA ranking exercises

Properties: the CA ranking approach was developed in mathematical psychology from conjoint measurement theory.[32] It is also rooted in Lancaster's theory of value.[31] The theoretical underpinning results in a strength of preference measure.[398]

Ref.	Empirical/methodological study	Acceptability to respondents	Cost	Internal consistency	Reproducibility	Validity
37	Chinburapa *et al.*, 1993 Determining physician's decision-making process in context of anti-infective drugs	51/63 physicians agreed to participate (offered $50 to do so). 3 were excluded: 2 made errors and 1 refused to complete the task. Average completion time was 35 minutes to rank 18 drug profiles				To increase validity respondents in the experimental group were told their decisions would be reviewed by peers; however, this did not influence responses Physicians shifted from using compensatory to non-compensatory decision-making as complexity of task increased
41	Orkin & Greenhow, 1978 Faculty members' evaluation of residents' clinical competence in terms of 6 attributes	34 faculty members; respondents stated task was difficult because it forced them to think about relative importance of criteria				Rankings correlated highly with rankings predicted on the basis of their sets of utilities (mean 0.93)
42	Parker & Srinivasan, 1976 Preferences for rural healthcare system	Door-to-door interviews yielded a response rate of 90.3% (177/196). 25 facility profiles were divided into "like" and "don't like" and then ranked. This process took 12–18 minutes			Results compared with a second set of profiles were not statistically significant at 0.10 level (n = 8 at 2 months after initial survey)	
43	Rosko *et al.*, 1983 Preferences for ambulatory care management	Convenience sample of 73 students. Ranked 18 profiles. They conducted a second ranking exercise where 2 attributes could not co-exist. Used orthogonal array to limit the scenarios to 27			In a comparative test, using 7 separate profiles, predicted ranks were within 1 rank of the actual observed in the main study	Coefficients for the levels of attributes (in terms of charges, travel time, office opening hours, waiting time, parking facilities, type of practice and sponsor) all behaved in line with *a priori* expectations. Authors note that the validity and reliability of CA have been well established in other disciplines but need assessing in healthcare

continued

CA ranking exercises *contd*

Properties: the CA ranking approach was developed in mathematical psychology from conjoint measurement theory.[32] It is also rooted in Lancaster's theory of value.[31] The theoretical underpinning results in a strength of preference measure.[398]

Ref.	Empirical/methodological study	Acceptability to respondents	Cost	Internal consistency	Reproducibility	Validity
21	Rosko & McKenna, 1983 Preferences for alternative healthcare plans (26 profile cards). Also, trade-off approach, which consists of paired comparisons of different levels of attributes	Convenience sample			More than 80% of ranks were within 1 rank of the actual profile when a second ranking of 7 profile cards was attempted. The trade-off approach yielded lower reproducibility or comparative validity	Coefficients for the levels of attributes (in terms of charges, travel time, office opening hours, waiting time, parking facilities, type of practice and sponsor) all behaved in line with *a priori* expectations
44	Shemwell & Yavas, 1997 Patients' preferences for a congregate care health facility					The coefficients for the attributes (cost, quality of nursing care, quantity of recreational facilities, quality of physical facilities, location, quality of rehabilitation programmes and staff attitudes) behaved as expected
46	Singh *et al.*, 1998 Preferences for growth augmentation therapy using CA	Consent rate for interviews was 83.7% (159/190). Pretest interviewees had little difficulty with the CA task		5 respondents gave 3 or more inconsistent answers, 11.95% gave 2 and 40.25% gave 1	Reliability coefficients of 0.7 and 0.66 were found for 10 respondents (after 7-month gap)	The coefficients for the attributes (magnitude of effect, certainty of effect, treatment route, costs, side-effects, child's attitude) behaved as expected. Other *a priori* hypotheses were supported: the risk-conscious group were most influenced by side-effects; the child-focused group were most influenced by the child's attitude; the cost-conscious group were most influenced by out-of-pocket costs; and the ease-of-use group were most influenced by route of treatment. In addition, the cost-conscious group were characterised by lower income and education, and the child-focused parents were characterised by moderate levels of income and education

Visual analogue scale (VAS)

Properties: the VAS has its theoretical base in psychometrics.[248,399–402] There is debate concerning whether the VAS measures strength of preference.

Ref.	Empirical/methodological study	Acceptability to respondents	Cost	Internal consistency	Reproducibility	Validity
253	Brazier *et al.*, 1999 Results of a systematic review of the use of health status measures in economic evaluation	VAS method widely seen as the most feasible and accepted of the health-state valuation techniques High response rates and high levels of completion[352,359,400,406–411] Gudex *et al.*[359] reports that of 3395 respondents only 3.2% of responses were excluded from analysis, whilst Silvertssen *et al.*[410] and Ferraz *et al.*[616] report completion rates of 95 and 100%, respectively The VAS method is cheaper and quicker than other methods (given that it can be carried out in a mailed questionnaire), but is not necessarily easier to complete[400,414]		Gudex *et al.*[359] cited 57.4% of respondents had no logical inconsistencies Mean overall inconsistency rate was 2.5%. Mean logical inconsistency at retest was 2.2%, slightly less than at test Dolan and Kind[360] found inconsistency rates of around 10% for first study, whilst every subsample of the second produced median rates below 3%	The following presents test–rest reliability results for the VAS: 1 week or less = 0.77[*]; 0.70–0.95[†] 4 weeks = 0.62[‡]; 0.89[§] 10 weeks = 0.78[#] 1 year = 0.49[¶] Correlations undertaken where specified: intraclass correlation coefficient (ICC)[†,‡,#]; Pearson correlation coefficient[§]; others unspecified Gudex *et al.*[359] mean ICC = 0.78. Only 13 respondents had an ICC < 0.6	Supporters of VAS argue the output is cardinal.[14,59,407,412] However, others challenge this.[61,62,246,399,617] Validity of VAS challenged on basis that preferences are elicited under certainty. However, Dyer and Sarin[426] suggest a measurable value function providing a link between such value and utility The VAS method may be susceptible to response-spreading,[51,407,425] further challenging the interval properties of the technique. Kaplan *et al.*[407] recognise the potential for response-spreading biases. However, they state that such biases may be controlled for by valuing 1 state at a time or through the use of a balanced design VAS methods were generally found to have a weak correlation with SG and TTO.[414,422,469,618] Currently, insufficient evidence to infer any correlation between the VAS or PTO. VAS methods correlate well with measures of health status (e.g. pain, clinical symptoms, etc.) compared with SG and TTO

[*] *O'Connor*[418]; [†] *Bakker* et al.[409]; [‡] *O'Brien and Viramontes*[419]; [§] *Gabriel* et al.[420]; [#] *Gudex* et al.[359]; [¶] *Torrance*[400]

Rating scales within CA

Properties: the CA rating approach has its theoretical basis in information integration theory and judgement analysis,[427,428] where again it has been argued that cardinal data can be obtained from individual responses to rating data. In addition, CA rating exercises are rooted in Lancaster's theory of value.[31]

Ref.	Empirical/methodological study	Acceptability to respondents	Cost	Internal consistency	Reproducibility	Validity
52	Chakraborty *et al.*, 1993 Patients' preferences for dental services using CA (10-point scale)	Incentive of $5 for completion. 3 practice profiles were used to get respondents used to the format. 42% response rate (126/300) with postal questionnaire and 1 follow-up				Authors note choices are easier for respondents and resemble real-life situations
53	Chinburapa & Larson, 1988 The importance of drug attributes in physician's prescribing	44% response rate (43/98) with postal questionnaire after 2 reminders; 12 profiles				
54	Diamond *et al.*, 1994 Specialty selection by fourth year medical students using CA	50% response rate (104/209) from 2 mailings and telephone follow-up				In terms of attributes, the authors note "if additional factors were added, or some ... were deleted, a different pattern of findings might emerge"
56	Graf *et al.*, 1993 Use of CA in design of an obstetrics unit (10-point scale)	Authors note CA can result in fatigue. Rating 16–24 profiles took around 10–20 minutes Telephone interviews difficult and the technique best suited to face-to-face interviews Mail surveys possible where incentives are offered				
58	Harwood *et al.*, 1994 Developing an outcome measure for handicap	Response rate 42% (101/240) and 79% of these (79) completed the interview Authors noted that interviews "were probably about as difficult as it is reasonable to undertake", though the response rate was comparable with other scaling methods			Reliability tested in a pilot of 9 respondents with a 2-week retest. Differences of within 1 category found. 5 'test' scenarios not used in the main study demonstrated good agreement with the main survey: Pearson 0.98 and Kendall 1.00	

continued

Rating scales within CA *contd*

Properties: the CA rating approach has its theoretical basis in information integration theory and judgement analysis,[427,428] where again it has been argued that cardinal data can be obtained from individual responses to rating data. In addition, CA rating exercises are rooted in Lancaster's theory of value.[31]

Ref.	Empirical/methodological study	Acceptability to respondents	Cost	Internal consistency	Reproducibility	Validity
59	Reardon & Pathak, 1990 Preferences for product attributes for antihistamine drugs	143/168 respondents provided useable responses				Noted that individuals will not be able to deal with a large number of attributes. 47 attributes identified and reduced to 7 by a panel of 10 patients
61	Ryan *et al.*, 1998 Preferences for patient health card (scale of 1–8)	67% response rate (67/100); 50 answered the CA question		Evidence of internal consistency of between 80% and 96% (with 'better' scenarios being rated more highly)		Evidence of internal validity (e.g. waiting had a negative sign, and better off people had higher opportunity cost of time). No evidence of ordering effects in terms of scenario ordering Within trial, only those respondents who had experience of a patient health card valued it
64	Wigton *et al.*, 1986 Physicians' weightings of types of clinical information used in diagnosing pulmonary embolism	Authors note that vignettes are convenient and easy to administer			3 individuals tested for reproducibility – found close clusterings of beta weights	Predicted answers for estimated model were correct 90% of the time

Schedule for the evaluation of individual quality of life (SEIQoL)

Properties: shares the same theoretical basis as rating scale CA (see pages 106–107).

Ref.	Study	Acceptability to respondents	Cost	Internal consistency*	Reproducibility	Validity†
70	McGee *et al.*, 1998 SEIQoL in healthy and gastroenterology populations	Correct completion by 100%, but by interview not self-completion		r = 0.74 for healthy respondents r = 0.54 for outpatients		R^2 = 0.75 for healthy respondents R^2 = 0.79 for outpatients
68	O'Boyle *et al.*, 1992 Measuring quality of life in hip replacement patients using SEIQoL	10/30 patients declined to take part or could not complete			Test–retest coefficient = –0.88	Differences between patients and controls; SEIQoL behaved similarly to standard scales (e.g. AIM)
67	Coen *et al.*, 1993 Use of SEIQoL to measure quality of life in mild dementia patients	Only 6 of 20 completed		r = 0.74 ‡ r = 0.75, r = 0.66		R^2 = 0.70 ‡ R^2 = 0.7, R^2 =0.72
69	Browne *et al.*, 1994 Use of SEIQoL to measure quality of life in the healthy elderly	10% non-completion (from 67 respondents) interviewed at home. Infers 90% able to understand		r = 0.74 for elicited and 0.69 for provided		R^2 = 0.75 for elicited and 0.79 for provided

* *Internal consistency measured using judgement reliability coefficient*

† *Internal (construct) validity has been assessed by assessing whether the variance in quality of life judgements is explained by the set of cues. This is measured by* R^2 *(which is a measure of the correlation between two variables)*

‡ *Unpublished results reported in Coen* et al.[67]

Likert scales

Properties: there is no constrained choice in Likert scales, meaning that respondents may be encouraged to overstate preferences. Debate exists on whether an ordinal or cardinal measure is estimated.

Ref.	Empirical/methodological study	Acceptability to respondents	Cost	Internal consistency	Reproducibility	Validity
80	Marks *et al.*, 1992 Quality of life in patients with asthma	283 questionnaires completed, reduced to 197 when excluding missing data 20-item questionnaire completed in less than 5 minutes 5-point Likert scale used because "easier to administer and interpret" than a VAS		Cronbach's α = 0.92, n = 77, outpatients Cronbach's α = 0.94, n = 87, community sample	Good test–retest reliability, intraclass correlation = 0.80 for 58 respondents	Weak correlation in expected direction for 3 medical indicators of asthma severity, supporting validity Whilst the 7-point Likert scale has been shown to be comparable to VAS, the 5-point Likert, however, has not been directly compared
83	Pfennings *et al.*, 1995 A comparison of Likert and VAS through medical students' opinion of preconditions of responsiveness	177 students responded, 9 included item non-response and were excluded from the analysis				Likert scores were transformed into 100-point VAS. In 3 of 9 items, mean was significantly higher than VAS, for the other 6 (3 of which were significant) VAS scores were greater. Authors conclude that their results provide modest support that VAS is preferable to Likert scales
77	Tymstra & Andela, 1993 Doctors' and nurses' opinions of healthcare policy, rationing and technology in The Netherlands. 5-point rating scale from 'very adequate' to 'very poor'	Response rate was 95.8%. Pilot response rates were 72.8% (of 471) nurses and 52.7% of doctors				

Semantic differential technique (SDT)

Properties: the technique offers no constrained choice, no strength of preference or any kind of theoretical basis.

Ref.	Empirical/methodological study	Acceptability to respondents	Cost	Internal consistency	Reproducibility	Validity
93	Girón & Gomezbeneyto, 1995 Study of family attitudes and relapse in schizophrenia	90 response rate with long-term follow-up (2 years)				Authors suggest "the semantic differential does not measure the relative's attitude but their accuracy in assessing ...", i.e. this particular technique does not measure what it should measure
89	Bowles, 1986 Attitudes towards the menopause	Highly acceptable		Cronbach's α = 0.96		
85	Holmes, 1974 Statistical evaluation of rating scales in marketing				Test–retest suggests that "the typical rating scales are not as reliable as we would like them to be"	Questionnaire-effect attitudes are modified somewhat from initial more extreme responses and within page bias: respondents prefer the left-hand side of page
429	Appels *et al.*, 1996 Self-rated health in Lithuania and The Netherlands	Data on large number of respondents was incomplete. Useful responses 41% and 68%				
91	Nichols *et al.*, 1996 Cancer detection and public education			Cronbach's α = 0.81		
88	Wilbur *et al.*, 1995 Study of attitudes etc towards the menopause	63% response rate		Cronbach's α = 0.96		
87	Valois & Godin, 1991 Methods paper on SDT			Cronbach's α for 4 areas such as smoking were: 0.67, 0.83, 0.49, 0.61		
90	Wikblad *et al.*, 1990 The patient's experience of diabetes and treatment	50 out of 62 (81%) possible respondents completed technique correctly		Cronbach's α = 0.96	High test–retest Reliability coefficient 0.93	

Guttman scales

Properties: Guttman scales have a theoretical basis in Facet Theory.[98] No information concerning constrained choice or strength of preference is provided.

Ref.	Empirical/methodological study	Acceptability to respondents	Cost	Internal consistency	Reproducibility	Validity
95	Kelloway & Barling, 1993 Studies of trade union activities of members in Canada	Low response rate (1 study had response rate < 10%)		*r* = 0.91 Coefficient of scalability = 0.60	Test–retest reliability (6 months had correlation of 0.70; $p < 0.01$	Scale positively related to measures of union loyalty, willingness to work for the union, and satisfaction from the union Consistent with the authors' hypotheses, extrinsic and intrinsic job satisfaction was negatively correlated with members' participation in union activities
98	Edmundson *et al.*, 1993 Utility of Guttman scale for technique development and validation	Convenience sample		*r* = 0.91		
99	Santos & Booth, 1996 Study meat avoidance among students	Only 66% of returned questionnaires were useful (no data on how any distributed) Response rate max. < 66%				
100	Sanner, 1994 Study of public attitude towards organ donation	Response rate of 65%				
102	Peterson, 1989 Study of dental health attitudes and knowledge	Response rate postal questionnaire: 71%		*Attitudes*: Coefficient of reproducibility = 0.91 Coefficient of scalability = 0.72 *Knowledge*: Coefficient of reproducibility = 0.82 Coefficient of scalability = 0.27		

r = *Coefficient of reproducibility*

Service quality (SERVQUAL)

Properties: is based in multiple discrepancy theory, and provides an ordinal measure of quality.

Ref.	Empirical/methodological study	Acceptability to respondents	Cost	Internal consistency	Reproducibility	Validity
167	Scardina, 1994 Patient satisfaction with nursing care; also, budget pie of 100 points divided between the 5 subscales to determine relative importance	Convenience sample of 10: 5 thought the survey adequate and 5 thought it too long. Reported no difficulty reading or understanding the instructions or statements		Established for 4 of the 5 attributes with Cronbach's α ranging from 0.74 to 0.98. Other attribute empathy perception (score) = 0.40		
168	Sargeant & Kaehler, 1998 Patients' satisfaction with GP services	182 personal interviews		One question was identical but negatively phrased. 66% were perfectly consistent and 90% were within 1 Likert point		
442	Oswald *et al.*, 1998 Hospital satisfaction – aim of determining whether dimensions other than those used by SERVQUAL were important	472 completed and useable responses to 660 postal surveys (72%)				Using a questionnaire with 13 items taken from Hospital Quality Trends*, the authors found that "analysis suggests the dimensions of perceived quality for the study sample differed from those measured by SERVQUAL"
171	Youssef *et al.*, 1996 Empirical study using SERVQUAL (on a 9-point scale) to measure patients' satisfaction with hospitals' quality of service	174 patients. Response rates were 29% for postal questionnaires: 80% for those handed out at hospitals and 36% when given out by GPs				The authors conclude that "The SERVQUAL instrument itself should be subject to continuous refinement as its application in the NHS grows"

* *Hospital Quality Trends is a trademark survey of hospital corporations of America*

continued

Service quality (SERVQUAL) *contd*

Properties: is based in multiple discrepancy theory, and provides an ordinal measure of quality.

Ref.	Empirical/methodological study	Acceptability to respondents	Cost	Internal consistency	Reproducibility	Validity
173	Headley & Miller, 1993 Satisfaction with medical services in a primary care setting	Of 967 initial mailings, 244 completed the 'pre-encounter' survey and 159 of these completed the 'post-encounter' survey		Cronbach's α = 0.87 (subscales ranged from 0.58 to 0.77)		Tentative convergent validity reported in comparison of SERVQUAL items and a universal quality judgement SERVQUAL appropriate for healthcare, but researchers should watch for situations that call for adaptation
176	Reidenbach & Sandifer-Smallwood, 1990 Modified SERVQUAL (5-point scale) to assess patient perception of service quality in inpatient, outpatient and emergency services	73% response rate from telephone survey to 300 patients		Cronbach's α for factors ranged from 0.83 to 0.96		
177	Lam, 1997 Methodological/empirical study using SERVQUAL to determine patients' perception of the quality of healthcare in Hong Kong	83/84 questionnaires were completed and only 1 of these was unusable		Cronbach's α of 0.64–0.88		5-point Likert scale rather than the recommended 7-point used since piloting indicated this would reduce frustration and increase response rate and quality of responses. It also led to the removal of negatively worded statements that had led to confusion and irritation for the respondents
178	Duffy *et al*., 1997 2-centre (USA and UK) empirical study of nursing home residents' evaluation of service quality	206 USA and 100 UK participants Authors note interviewing took 1–1.5 hours		Cronbach's α = 0.95 and 0.88 respectively, although 2 subscale ratings in the UK sample were ~ 0.4	See 'Validity'	Reliability and validity of SERVQUAL has been established in non-health literature

continued

Service quality (SERVQUAL) *contd*

Properties: is based in multiple discrepancy theory, and provides an ordinal measure of quality.

Ref.	Empirical/methodological study	Acceptability to respondents	Cost	Internal consistency	Reproducibility	Validity
180	Mitchell *et al*., 1999 Using modified SERVQUAL (5-point scale) to assess employees' satisfaction with the quality of nursing services. Minor wording modifications	43% response rate (86/200)		Cronbach's α = 0.81–0.96 for the 5 subitems		Content validity was established through qualitative studies
370	Babakus & Mangold, 1992 Evaluating the application of SERVQUAL to hospital services	A 5-point Likert scale was used instead of the original 7-point. 22% response rate from a postal questionnaire (443/1999)		Cronbach's α for subscales ranged from 0.759 to 0.903		Evidence of correlation between expectations and perceptions
182	Duffy & Ketchand, 1998 Use of SERVQUAL to assess patient satisfaction with nursing home care	206 nursing home residents (mean age 79) were included; no mention of response rate		Cronbach's α = 0.95		Unexplained variation (55%) infers that dimensions which may have influenced satisfaction were excluded
447	Hart, 1996 Empirical/methodological study reporting on 4 studies on outpatient clinics which have used SERVQUAL to measure quality. 5 dimensions weighted to sum 100					7-point Likert scale may not have equal intervals and respondents may "actually deploy an 'increasing resistance' model in which it is easier (in psychometric terms) to move from the central point (point 4) to its immediate neighbour (point 5) than it is to move from a position of near perfect satisfaction (point 6) to perfect satisfaction (point 7)". Instead, the author suggests assigning cardinal values to each point as follows: point on scale 1 (value –6); 2 (–3); 3 (–1); 4 (0); 5 (1); 6 (3); 7 (6)
445	Tomes & Ng, 1995 Development of a questionnaire, based on SERVQUAL, to measure hospital service quality			The 49-item scale had good internal consistency (0.959); individual dimensions also revealed high consistencies		Dimensions were not taken from SERVQUAL but were developed after hospital staff interviews

Simple choice exercises and random paired comparisons

Ref.	Empirical/methodological study	Acceptability to respondents	Cost	Internal consistency	Reproducibility	Validity
184	Lewis & Charny, 1999 Subset of the Cardiff study[183] dealing with the relative value of life dependent on age comparisons only					Respondents favoured the 5-year-old over the 70-year-old by a ratio of 84:1 and the 35-year-old to the 60-year-old by 14:1, but favoured the 8-year-old over the 2-year-old by 5:3 (see below). Highlights complexity of age-based choices
183	Charny *et al.*, 1989 Cardiff study eliciting preferences of 1 specific life relative to another. Respondents had to choose between 2 hypothetical individuals who differed by only 1 characteristic (age, marital status, gender, smoking/drinking status, and employment)	Of the 722 respondents, 220 (31%) made all 13 choices, whilst nearly 75% made choices in half or more of the 13 questions. The fact that 69% were unable to make all the choices, indicates, in the opinion of the authors, that respondents took the exercise very seriously				As in earlier findings, results favoured young over old, married over single, women over men and non-smokers and drinkers over smokers and drinkers. The instance where an 8-year-old child was favoured over a 2-year-old was explained by the greater parental investment in terms of effort and emotion
186	Ryynanen *et al.*, 1996 Methodological/empirical study of prioritisation between a number of 'ethical value indicators' (consisting of age, income, severity of disease, prognosis, social status, cost of treatment, and origin of disease)	Authors note that members of the public (n = 49) completed the questionnaires quickly; however, 'most' of the public completing the exercise found difficulty in making choices between 2 alternatives. Some found that the questionnaire made them anxious and 1 reacted aggressively. The medical and nursing undergraduates (n = 104) completed the questions quickly and without difficulty			Test–retest reliability was measured by 8 respondents answering 3 different sets of RPS questions. Authors state that reliability was "good"	Criterion validity was examined by comparing the undergraduate students' responses to the RPS questions compared with conventional questionnaires, and comparing a non-structured interview with the RPS for members of the public; the results were comparable

CA choice-based questions

Properties: choice-based CA is rooted in random utility theory[60,61,193,198,201,203] and Lancaster's theory of value.

Ref.	Empirical/methodological study	Acceptability to respondents	Cost	Internal consistency	Reproducibility	Validity
187	Bryan *et al.*, 1998 Preferences for use of magnetic resonance imaging in knee injuries	Convenience sample of 134 students: all responded				Theoretical validity was supported with *a priori* hypothesis confirmed 82 (61%) respondents were not willing to trade
188	Chakraborty *et al.*, 1994 Empirical study using CA to evaluate health insurance programmes (4 alternatives, each with 24 attributes)	662 respondents participated from 777 approached. Each was paid $15 for participation. Postal questionnaire				
189	Farrar & Ryan, 1999 Consultant preferences for alternative clinical service developments					Theoretical validity was supported with *a priori* hypothesis confirmed The authors found no evidence of ordering effects in terms of ordering of attributes within scenarios
191	Ferguson *et al.*, 1994 Pilot study of elderly patients' attitudes over choice of drugs for treatment of hypertension: 3 attributes presented in pairs only	44 patients, mean age 75 years. 21/24 agreed to participate in a clinic group and 25/28 in an office group. Response rate = 88%. 2 patients had difficulty understanding the questionnaire and were excluded				
619	Freeman *et al.*, 1998 Student preferences for healthcare needs using CA (maximum difference CA method)	The authors note the method is simpler than ranking and rating CA methods and respondent fatigue is reduced. Also, level of accuracy is increased				
192	Hakim & Pathak, 1999 Comparison of Euroqol (European quality of life) health-state preferences using choice-based CA, rating scales and SG					Evidence of convergent validity between choice-based CA, rating scales and SG

continued

CA choice-based questions *contd*

Properties: choice-based CA is rooted in random utility theory[60,61,193,198,201,203] and Lancaster's theory of value.

Ref.	Empirical/methodological study	Acceptability to respondents	Cost	Internal consistency	Reproducibility	Validity
193	McIntosh & Ryan, 1999 Preferences for location of treatment	Response rate was 56% (556/1000) for a postal questionnaire		6% of respondents provided intransitive responses		Theoretical validity was supported with *a priori* hypothesis confirmed 63% of the sample violated the continuity axiom, i.e. they showed dominant preferences
195	Propper, 1990 Estimation of the value of reducing waiting time					Theoretical validity was supported with *a priori* hypothesis confirmed The author notes the presence of non-trading (or lexicographic) preferences in 4%
196	Ratcliffe & Buxton, 1999 Patient preferences for outcome and process attributes in liver transplantation	Response rate was 89% (189/213) using postal questionnaire, including covering letter from physician and with 1 reminder. 59% (111) found the questionnaire was not difficult to complete and 26% (48) thought it slightly difficult. Mean time for completion of the questionnaire was 16 minutes (range 10–60)		9% of respondents gave inconsistent answers		Study used levels of attributes that most closely resembled existing services on offer. Internal validity confirmed
197	Ryan, 1999 Empirical/methodological application of CA to IVF treatment	331/414 questionnaires were completed: 59 non-responses		Low levels of inconsistency: first questionnaire = 4 and second = 6		Theoretical validity was supported with *a priori* hypothesis confirmed The author found support for internal validity in terms of *a priori* expectations All respondents were traders
205	Ryan & Farrar, 2000 Application of CA to orthodontic services (location, waiting time): 5-point scale	157/160 questionnaires were completed	Authors note the technique was well received by policy-makers	8 respondents answered at least 1 of choices 2 or 13 inconsistently and 37 answered choice 10 inconsistently		Theoretical validity was supported with *a priori* hypothesis confirmed

continued

CA choice-based questions *contd*

Properties: choice-based CA is rooted in random utility theory[60,61,193,198,201,203] and Lancaster's theory of value.

Ref.	Empirical/methodological study	Acceptability to respondents	Cost	Internal consistency	Reproducibility	Validity
199	Ryan & Hughes, 1997 Women's preferences for miscarriage management	33% response rate (196). 85 respondents thought it easy or extremely easy, 45 difficult and 9 extremely difficult		Respondents gave only a small number of inconsistent answers		Evidence was found to support theoretical validity 32 respondents consistently had a preference for 1 type of treatment over another, and were not willing to trade
200	Ryan & Wordsworth, 2000 Women's preferences for cervical screening programmes	A pilot found 13 choices to be too many 641 usable responses were received from 2000 after 2 reminders				Support for theoretical validity Levels of attributes were varied, whilst the coefficients were not significantly different across 5 of the 6 attributes; mean WTP was significantly different for 4 of the 5 welfare estimates
201	San Miguel *et al.*, 2001 Preferences for of menorrhagia	Convenience sample of 146 women. No reminders. Response rate was 51% (75/146). Respondents found questionnaire not too difficult, although there was a spread of reported difficulty. Time for completion was 14 minutes on average				Support for theoretical validity Evidence of limited trading, with 36 (48%) respondents consistently having a preference for 1 type of treatment over another
357	San Miguel *et al.*, 1999 Testing for completeness and stability preferences for out-of-hours care	54% response rate (735/1364): 731 were usable			Kappa coefficient of 0.69 and 0.68	
202	Scott & Vick, 1999 Patient preferences for attributes of the doctor–patient relationship	18.4% response rate (734/3983) postal questionnaires (no reminders). 82% found the questionnaire okay or easy to complete whilst 18% found it difficult or extremely difficult. Authors note that due to the low response rate, those returning the questionnaire may have been the ones who found it easy		97% of responses consistent		Prior hypotheses were confirmed, indicating theoretical validity (i.e. patients preferred more information to less, prefer easy explanations from the doctor, prefer the doctor listening to what they have to say, and prefer a shorter waiting time)

continued

CA choice-based questions *contd*

Properties: choice-based CA is rooted in random utility theory[60,61,193,198,201,203] and Lancaster's theory of value.

Ref.	Empirical/methodological study	Acceptability to respondents	Cost	Internal consistency	Reproducibility	Validity
206	Spoth & Redmond, 1993 Patient preferences for prevention programmes	Use ACA (adaptive CA) software to facilitate 202 telephone interviews (99.5% response rate). Of a 25-minute interview, the CA part took around 15 minutes				0.90 or greater correlation for 86% of the sample between CA results and rating exercise Method is similar to that faced in real-life situations where respondents choose between different programme profiles
620	Szeinbach *et al.*, 1999 Eliciting patient preferences for health states using CA and VAS					CA and VAS results were comparable, showing convergent validity Similar responses for VAS and CA in 2 profiles with similar outcomes but worded differently
203	Vick & Scott, 1998 Relative importance to patients of the attributes of GP consultation	63% response rate (101/160); mean time of completion about 10 minutes; 80% of respondents found the questionnaire okay to very easy, whilst 19% found it quite difficult and 2% very difficult		For the 1 dominant option, 80–90% of responses were consistent		Theoretical validity is supported for 2 of the 3 attributes, but cost was not significant Some evidence of ordering effects

Analytic hierarchy process (AHP)

Properties: Saaty's[219] axioms underlying the theoretical foundation of the AHP are noted[227,232,233] and the method reflects intensity of preference.[238]

Ref.	Empirical/methodological study	Acceptability to respondents	Cost	Internal consistency	Reproducibility	Validity
225	Schwartz & Oren, 1988 Patient preferences for a new intervention for recovery from myocardial infarction	83% completed the questions. 28 pairwise comparisons not seen as excessive, and were easily understood		5% of respondents were excluded from the analysis because their consistency ratios exceeded 0.50 (the others ranged from 0.11 to 0.23)		
235	Dolan, 1995 Study seeking to determine whether patients are willing and able to use the AHP, with choices of alternative screening regimens for colon cancer used as an example	20 patients participated, 18 (90%) were capable (judged as able to complete the exercise in ≤ 45 minutes) and willing to go through AHP analysis before making a clinical decision Patients completed the AHP exercise on a laptop computer. Following the AHP exercise, respondents were asked evaluation questions: 15% said the exercise was hard to understand; 95% learned useful information during the exercise; 90–100% said they would like AHP to influence decision-making				Findings are consistent with previous results,[232] although this study involved small numbers Validity of AHP has contested through rank reversal (in which a change in relative desirability is caused by the introduction of another alternative into the analysis). New method of AHP precludes rank reversal
226	Dolan, 1989 Determine best antibiotic regimen for woman with acute pyelonephritis					Study demonstrates the usefulness of AHP, which led to a change in prescribing behaviour to improve the process of patient care
239	Javalgi *et al*., 1989 Presentation of the AHP and illustration applied to public preferences for hospital attributes when selecting a hospital for treatment	47% (235) response rate from 500 postal questionnaires (with a $1 incentive); a further 15 of which were unusable				

continued

Analytic hierarchy process (AHP) *contd*

Properties: Saaty's[219] axioms underlying the theoretical foundation of the AHP are noted[227,232,233] and the method reflects intensity of preference.[238]

Ref.	Empirical/methodological study	Acceptability to respondents	Cost	Internal consistency	Reproducibility	Validity
236	Dolan, 1993 Illustration of the AHP for treatment options for a man with a common serious disease	Previous work has demonstrated feasibility with patients,[453,455] 80% of the former[453] rating the AHP feasible and 22 patients and 25/26 physicians in the latter[455] successfully completing the exercise within a reasonable amount of time				Author notes that this and previous applications[218,227,228,237,621] demonstrate the AHP's suitability for medical decision-making and counteract questioning by Eckman[622]
621	McNeil *et al*., 1982 Patient, student and physician choices of alternative therapies for lung cancer using variations of background information provided	Respondents appeared to understand the data presented, although they were influenced by the nature of the data and the form in which it was presented				
229	Dolan & Bordley, 1994 Use of AHP concerning isoniazid prophylaxis in positive tuberculin tests					Validity of AHP is debated concerning the rank reversal phenomenon. Authors used a method of aggregation where rank reversal cannot occur
452	Peralta-Carcelen *et al*., 1997 Empirical study of the preferences of pregnant women, physicians and paediatricians for reducing incidence of neonatal group B streptococcal sepsis using AHP	90% (83/92) of pregnant women, 51% (40/78) of obstetricians and 67% (40/60) of paediatricians participated. 89% (74) of women and 90% (72) of physicians understood the interview format. A 'few' of the women patients had difficulty understanding some of the criteria, as did some of the physicians. 87% of women patients, 88% of paediatricians and 75% of obstetricians liked the interview format. Concern expressed that women with limited educational backgrounds might have difficulty understanding the model; however, no evidence of this. Given the modest educational background of the sample, the authors believe the method to be generalisable		3 of the women patients provided inconsistent responses whilst there were no inconsistencies with the physicians		

Standard gamble (SG)

Properties: SG is rooted in EUT[245,352,456,457] and is held to measure strength of preference.[458]

Ref.	Empirical/methodological study	Acceptability to respondents	Cost	Internal consistency	Reproducibility	Validity
253	Brazier *et al.*, 1999 Results of a systematic review of the use of health status measures in economic evaluation	SG Empirical studies using SG have reported response and completion rates to demonstrate a high degree of acceptance. Many studies have reported completion rates between 80% and 100%[361,411–413,459–462] The author cites that "although SG has shown some completion problems within particular studies, these have been no worse than similar difficulties associated with other instruments used at the same time". For example, where completion problems occurred in studies by Patrick *et al.*[463] and van der Donk *et al.*,[414] which used SG, TTO and VAS methods, SG was not seen to be more burdensome than other methods employed Llewellyn-Thomas *et al.*[361] state that the SG is a complex method	Given that SG questions are collected by trained interviewers, the technique will be relatively expensive[352,400]	Llewellyn-Thomas *et al*[361] report 54 of 64 (84%) respondents ranked 5 health states with *a priori* expectations Lenert *et al.*[411] found 78% of healthy subjects and 44% of patients had a consistent rank ordering of preferences among VAS and pairwise comparison, whilst only 80% and 61%, respectively, had a consistent rank ordering of preferences among SG and pairwise comparison ratings	Froberg and Kane[352] present evidence of good inter-rater reliability (r = 0.77; from Torrance[400] and test–retest reliability (r = 0.80)) The following details more recent studies identified by Brazier *et al.*: 1 week or less = 0.8[*]; 0.77–0.79[†] 4 weeks: 0.82[‡] 6–16 weeks: 0.63 props[§]; 0.74 no props[§] 1 year: 0.53[#] other: 0.82[¶] time unspecified: 0.80[**] Intraclass correlation coefficient[†,‡,**]; Pearson correlation coefficient[§]; others unspecified	SG based on EUT developed by von Neumann and Morgenstern[††]. Given this, it is often viewed as the classic method of decision-making under uncertainty.[245] However, there is a substantial body of evidence showing that individuals consistently violate the axioms of EUT SG valuations may be affected by differing attitudes to risk, i.e. whether people are risk-averse, risk-neutral or risk-seeking. Loomes and McKenzie[623] suggest that at times people may be a mixture of all 3 Evidence supports the convergent validity of SG with TTO[462,467,468,469] and WTP[80,324] SG does not correlate well with either health status or VAS.[409,422] SG values have generally been found to exceed those of VAS methods[409,412,413,414,419,459,469]

[*] *O'Connor*[418]; [†] *Bakker* et al.[409]; [‡] *O'Brien and Viramontes*[419]; [§] *Dolan* et al.[462]; [#] *Torrance*[400]; [¶] *Reed* et al.[466]; [**] *Gage* et al.[459]

[††] *Von Neumann J, Morgenstern O. Theory of games and economic behavior. Princeton, NJ: Princeton University Press; 1944*

Time trade-off (TTO)

Properties: given that the TTO technique involves trading, it can be argued that it is rooted in consumer theory. As with the SG, the TTO method inherently provides the respondent with a constrained choice that elicits utility numbers that are held to represent strength of preference.

Ref.	Empirical/methodological study	Acceptability to respondents	Cost	Internal consistency	Reproducibility	Validity
253	Brazier *et al.*, 1999 Results of a systematic review of the use of health status measures in economic evaluation	TTO This technique has proved to be a practical and acceptable method of health-state valuations in a wide variety of studies, achieving completion rates between 90% and 100%[58,345,363,463,484–488,491] Argued to be easier than SG[62]	Given that SG questions are collected by trained inter-viewers, the technique will be relatively expensive[352,400]	Dolan *et al.*,[362] in a large general population sample of 3395, found the TTO method to be consistent Laupacis *et al.*[624] found similar results Ashby *et al.*[363] compared TTO scores with the rank ordering of states given by respondents. The results show a considerable degree of consistency in ranking and that rank ordering was consistently reflected in the mean TTO values	1 week or less: 0.87[*] 4 weeks: 0.81[†]; 0.63[‡] 3–6 weeks: 0.5–0.75[§] 6 weeks: 0.63–0.80[†]; 0.85[#] 10 weeks: 0.73[¶] 6–16 weeks: 0.83[**] (props); 0.55[**] (no props) 1 year: 0.62[††] Correlations undertaken where specified: intraclass coefficient[**]; Pearson correlation coefficient[‡,††]; others unspecified	Whilst TTO is not directly related to any specific theory, its sacrifice element links it to consumer theory[362,404] Buckingham *et al.*[625] align the TTO method with the welfare economic approach of 'compensating variation', where welfare gain is measured by compensating loss of something else that is valuable so that the respondent is returned to the original level of welfare Interval properties of technique challenged, as has the realism of the choices posed[62] TTO challenged due to the certainty context of questions posed[626] Empirical evidence supporting time preference effects[502,627] and the violation of the assumption of proportional TTO[362,497–499] challenge the validity of TTO Empirical studies highlight the unwillingness of individuals to trade life expectancy[362,484,490,494,501]

[*] *O'Connor*[418]; [†] *Churchill* et al.[491]; [‡] *Gabriel* et al.[420]; [§] *Ashby* et al.[363]; [#] *Molzahn* et al.[492]; [¶] *Dolan* et al.[362]; [**] *Dolan* et al.[462]; [††] *Torrance*[400]

Person trade-off (PTO)

Properties: although the technique is seen as intuitively appealing,[251] it has been criticised for a lack of any theoretical basis,[250,253,503] other than psychometric qualities surrounding adjustment or equivalent stimuli.[248]

Ref.	Empirical/methodological study	Acceptability to respondents	Cost	Internal consistency	Reproducibility	Validity
253	Brazier *et al.*, 1999 Results of a systematic review of the use of health status measures in economic evaluation	Acceptability of PTO currently unknown. However, given low response rates and nature of questions posed, data collection will require interviews[248,251,364,504] Respondents may find questions difficult given that they have to trade lives. Nord found that 17 out of 53 subjects showed an unwillingness to take part[251]		Ubel *et al.*[364] note problems with consistency of responses. 11/53 people excluded from study due to failing to answer the rationing scenario (PTO) consistently, although this study was self-administered	Patrick *et al.*[248] report a comparatively low correlation (Pearson; $r = 0.60$) with respect to inter-rater reliability, compared with the VAS ($r = 0.75$–0.77) Nord[628] reports that retest findings from 20 individual PTO responses ("some weeks after the first response") showed a mean difference of 40%	Although the technique is intuitively appealing,[251] the theoretical basis is still to be developed The technique has been argued to possess interval properties[57,61] and has been said to ask the right question (i.e. trade-offs between people)[423] PTO may be sensitive to framing effects (the arguments mentioned in the questions, the start-point, the numbers in the pairwise comparison, and the choice of decision context[251] There is some evidence of internal validity, with PTO responses consistently reflecting a preference to treat those in a worse state.[364,423] SG and 3 versions of PTO directly ordered the health states in line with the expected ordering. The VAS did not[423]

continued

Person trade-off (PTO) *contd*

Properties: although the technique is seen as intuitively appealing,[251] it has been criticised for a lack of any theoretical basis,[250,253,503] other than psychometric qualities surrounding adjustment or equivalent stimuli.[248]

Additional articles concerning health status measures not reviewed in Brazier *et al.*[253*]

Ref.	Empirical/methodological study	Acceptability to respondents	Cost	Internal consistency	Reproducibility	Validity
464	Hall *et al.*, 1992 A cost–utility analysis of mammography screening	In comparing SG, TTO and VA, SG was found to be the most difficult technique				
416	Shiell *et al.*, 1993 The consistency of VAS and TTO techniques for eliciting preferences	Problems were noted with both the TTO and VAS				23 (30%) patients gave similar responses to the 2 valuation methods For the majority of patients (45; 59%), the 2 resulted in substantially different values. Of these, 7 (9%) patients selected their VAS response in preference to the TTO and 8 (11%) patients selected their TTO response in preference to the VAS. The remaining 30 patients stuck with their original responses despite the substantial difference in the responses given and the pressure to be consistent
365	Badia *et al.*, 1999 To assess inconsistent responses in preference-elicitation methods for health states	294 (98%) questionnaires used in the final analysis		% of respondents with logical inconsistencies: 26% for VAS; 59% for TTO. Mean rate: 1.14 for VAS; 3.95 for TTO. Increasing age and lower levels of education positively associated with inconsistent responses for both VAS and TTO		

continued

Person trade-off (PTO) *contd*

Properties: although the technique is seen as intuitively appealing,[251] it has been criticised for a lack of any theoretical basis,[250,253,503] other than psychometric qualities surrounding adjustment or equivalent stimuli.[248]

Additional articles concerning health status measures not reviewed in Brazier *et al.*[253*] *contd*

Ref.	Empirical/methodological study	Acceptability to respondents	Cost	Internal consistency	Reproducibility	Validity
494	Robinson *et al*., 1997 Valuing health sates using VAS and TTO					29/43 respondents ranked and scored at least 1 state above death on the VAS but subsequently rated that state as worse than dead in the TTO. When asked to explain their decision, all 29 confirmed that they would rather die immediately than spend 10 years in the given health state. When asked whether their VAS answer meant that they personally preferred 10 years in that health state to immediate death, 12/29 said that it did, and 14 said that it did not, whilst 3 did not know 15/43 respondents (34.8%) refused to trade-off even a few days or weeks in order to avoid a health state that they had placed below 11111 on the VAS. All 15 confirmed that their VAS response meant that they considered 10 years in that state to be worse than 10 years in 11111. They did not, however, seem to translate this into a willingness to trade-off time to avoid that state. Only 1/15 did not trade-off any time at all throughout the TTO exercise, apparently objecting to the task on religious grounds. Thus, the authors state that it does not appear that refusing to 'play the game' in the TTO would, in itself, account for this disparity. Rather, the predominant message was that, as long as they could cope with the state in question, they would not consider giving up any of the 10 years to avoid it The authors cite that TTO states had a 'threshold tolerability' below which states have to fall for respondents to be willing to trade any time at all. Thus, TTO may not be suitable for use in certain clinical settings Younger respondents (in TTO) find the worse than dead scenario less plausible than older respondents

[*] *Other studies identified, not reviewed in Brazier* et al., *are mentioned in choice-based CA*[192,620] *and WTP*[417,456,470,496]

continued

Person trade-off (PTO) *contd*

Properties: although the technique is seen as intuitively appealing,[251] it has been criticised for a lack of any theoretical basis,[250,253,503] other than psychometric qualities surrounding adjustment or equivalent stimuli.[248]

Additional articles concerning health status measures not reviewed in Brazier *et al.*[253*] *contd*

Ref.	Empirical/methodological study	Acceptability to respondents	Cost	Internal consistency	Reproducibility	Validity
458	Shackley & Cairns, 1996 Evaluating the benefits of antenatal screening using SG	52 eligible women were interviewed: 3 were excluded from analysis because they would not have a diagnostic test under any circumstances				Evidence of internal validity with 41/49 (84%) responders willing to accept an increase in the risk of fetal loss as their risk of cystic fibrosis increases

[*] *Other studies identified, not reviewed in Brazier* et al., *are mentioned in choice-based CA*[192,620] *and WTP*[417,456,470,496]

Willingness to pay (WTP)

Properties: has its theoretical basis in welfare economic theory[505] and is held to measure strength of preference.

Ref.	Empirical/methodological study	Acceptability to respondents	Cost	Internal consistency	Reproducibility	Validity
257	Klose, 1999 Review of methodological aspects in the use of WTP in healthcare	Evidence of non-response and protest answers with the OE method			Reproducibility rarely investigated[419,508,519,527] and still to be proven[516]	Evidence of internal validity through positive influence of income (and in several cases social class) and health gain in many WTP studies Evidence of convergent validity through correlation with other measures of health benefits found in some but not all studies[419,495,520] Evidence of strategic behaviour[532,544] Evidence of 'yea-saying' with CE method,[356] range bias with PC[355,506] and starting point bias with bidding game[353,356,552,558]
532	Anderson *et al.*, 1997 Empirical study on cataract surgery waiting list patients' WTP for shorter waiting times CE method	464 responses to the WTP question, by telephone; no data on response rate				Some evidence of theoretical validity, though income was not statistically significant WTP compared with actual behaviour (since respondents were able to have the surgery at private clinics). From the 3 centres, 25%, 15% and 12% said they were willing to pay the market price to reduce waiting by 1 month; however, only 1.7% actually did so
530	Appel *et al.*, 1990 WTP for reduced risk of low osmolality contrast media. Uses bidding game approach	95 (out of 100) outpatients were able to complete the questionnaire				Some evidence of internal validity, though some results were opposite to *a priori* expectations No respondents were willing to pay more for a reduced risk of the minor side-effects than for a reduced risk of both major and minor side-effects. However, 12% were willing to pay more for a reduced risk in 1 of the 8 side-effects than a reduction in all 8, and 22% were willing to pay more for a reduced risk in 1 of the minor side-effects than for a reduced risk in all 4 minor side-effects

continued

Willingness to pay (WTP) *contd*

Properties: has its theoretical basis in welfare economic theory[505] and is held to measure strength of preference.

Ref.	Empirical/methodological study	Acceptability to respondents	Cost	Internal consistency	Reproducibility	Validity
470	Bala *et al.*, 1998 Comparison of WTP and SG for valuing health outcomes associated with shingles Double-bounded CE method	Focus groups revealed both SG and WTP questions to be straightforward		No difference in consistency between WTP and SG		Respondents gave meaningful answers for both SG and WTP No signifcant relationship between SG and WTP, questioning whether the 2 approaches are equivalent preference-based outcome measures
549	Baron, 1997 WTP for relative and absolute risk reductions OE method	95 (from 108) students completed the first questionnaire; 29 completed the second				*Experiment 1* WTP unresponsive to quantity of lives saved. Mean WTP $143 for 900 lives saved and $74 for 90. Concluded that respondents were influenced by proportion of lives saved (measured against total number at risk) as well as the number saved *Experiment 2* Correlation ranging from 0.8 to 1 between WTP for a percentage and WTP for number reduction in death in 26% of respondents. This indicates respondents insensitive to quantity
456	Blumenschein & Johannesson, 1998 Relationship between quality of life techniques (rating scales, SG, TTO) and WTP Used both bidding and CE WTP approaches	69 patients completed the questionnaire at interview. No one refused to answer questions or gave obvious 'protest' answers				Rating scales gave lowest utility measure and SG highest No significant difference between bidding and CE WTP, although the CE mean was higher

continued

Willingness to pay (WTP) *contd*

Properties: has its theoretical basis in welfare economic theory[505] and is held to measure strength of preference.

Ref.	Empirical/methodological study	Acceptability to respondents	Cost	Internal consistency	Reproducibility	Validity
356	Chestnut *et al.*, 1996 WTP for changes in angina symptoms Triple-bounded CE question followed by OE question: respondents completed both	50 men with a history of chest pain: 2 refused CE variant and 1 refused OE				5 respondents gave a lower response to the OE question than they indicated 'yes' to in the CE question. Results for the 2 methods gave consistent answers – not surprising since the same respondents answered both questions Evidence of 'yea-saying' in CE responses, starting point bias in OE WTP not sensitive to scope effects – mean responses not significantly different for varying amounts of angina prevention. Some respondents indicated they were concentrating on what they could afford to pay rather than the quantity of the intervention
355	Donaldson *et al.*, 1997 WTP for cystic fibrosis carrier screening OE method	51% (of 450) response rate. Authors note that given the low response rate, conducting a postal WTP study outside a trial may be 'problematic' 75% of respondents found the WTP questions difficult to answer		23% of responses inconsistent – stated a preference for method 1, or indicated no preference, but gave a higher WTP for method 2		Social class significantly associated with WTP for method 1. The only significant variable for method 2 was being a single mother OE questions may not force respondents to give their maximum WTP, but rather to indicate an acceptable amount Mean WTP similar to that of earlier study.[629] However, this may be that estimates of NHS costs were similar rather than WTP (since inconsistent responses were explained by cost-based responses)
512	Donaldson *et al.,* 1997 Comparing OE and PC approaches	Higher response and completion rate with PC approach (65% for PC, 61% for OE, not statistically significant). 94% of respondents answered the WTP question in the PC questionnaire, 84% in the OE questionnaire. No zero responses in PC, 6 in OE				Stronger association with ability to pay (using the proxy of social class) with PC approach. Mean WTP progressively fell as moved from social class I to V Mean and median WTP higher in PC approach, although only statistically significant at 0.1% level

continued

Willingness to pay (WTP) *contd*

Properties: has its theoretical basis in welfare economic theory[505] and is held to measure strength of preference.

Ref.	Empirical/methodological study	Acceptability to respondents	Cost	Internal consistency	Reproducibility	Validity
354	Donaldson, 1999 Report from European Commission study to evaluate method of WTP. Incorporates 2250 respondents across 6 European countries	Findings indicate the method was feasible, with only 1 of the 6 countries showing > 10% protest answers to the WTP question			Good test–retest reliability	WTP insensitive to scope effects and ordering effects PC did not demonstrate range bias when additional bids were added Insurance-based questions led to less zero responses and higher WTP than community-based questions CE method results in substantially higher estimates compared with the PC method The incremental or marginal approach gives more consistent response with stated rankings The provision of additional information led to higher WTP values
521	Dranitsaris, 1997 Feasibility of WTP as a measure of supportive cancer care OE approach	Random sample of 50 patients receiving chemotherapy. WTP technique simple to administer and easy for respondents to understand				Family income, marital status and residence were all significantly associated with WTP WTP order consistent with rating scale rank order of adverse effects
630	Easthaugh, 1991 Value of risk-free blood	Respondents were 20 blood distribution managers and 50 health service administration graduate students				Authors found diminishing WTP with increasing marginal benefit. For a reduction in risk of 2 units (to zero risk) median WTP was $4 for managers and $3 for graduates. For a reduction from 4 units to 2 units it was $2 for both. For a reduction of 6 units (9 to 3) WTP was $5 and $4. A 10-unit reduction (20 to 10) revealed WTP of $1 for both groups

continued

Willingness to pay (WTP) *contd*

Properties: has its theoretical basis in welfare economic theory[505] and is held to measure strength of preference.

Ref.	Empirical/methodological study	Acceptability to respondents	Cost	Internal consistency	Reproducibility	Validity
519	Flowers *et al*., 1997 Computer-based assessment of WTP via bidding-type method	52 "young and well-educated" volunteers paid $10 to take part. All completed questionnaire: 71% indicated they understood the method, 79% thought it was very clear, 84% thought it useful for moderately difficult decisions, and 61% were comfortable or very comfortable using WTP to make healthcare decisions The study indicated the viability of automated questionnaires on computer				Good test–retest reliability (0.796) at 2 weeks
367	Frew *et al*., 1999 WTP for 2 alternative tests for colorectal cancer screening OE/PC methods comparison	Partial analysis so far revealed higher response rate with PC (85% and 77%) than OE (70% and 66%), although not statistically significant		20% of respondents who preferred the first test or expressed no preference, gave a higher WTP for the second test		Neither PC nor OE WTP estimates were significantly associated with social class No statistical difference between mean WTP for OE and PC methods
546	Garbacz & Thayer, 1983 Valuing a senior companion programme (WTP and WTA)					Insignificant difference between bid for a 25% and 75% reduction in the service Equivalent variation bids were around half of the compensating variation
543	Granberg *et al*., 1995 Value of IVF services using OE approach					55% of couples answering the WTP question were willing to pay more than £10,000 for IVF treatment, which equates with true cost of £9410
506	Johannesson *et al*., 1991 Cost–benefit analysis of non-pharmacological treatment of hypertension, incorporating WTP for benefit assessment 2 PC methods	Response rate was 99% for PC1 and 97% for PC2				The difference between the mean of the 2 PC methods not statistically significant

continued

Willingness to pay (WTP) *contd*

Properties: has its theoretical basis in welfare economic theory[505] and is held to measure strength of preference.

Ref.	Empirical/methodological study	Acceptability to respondents	Cost	Internal consistency	Reproducibility	Validity
263	Johannesson *et al.*, 1993 WTP for antihyperintensive therapy Used CE question with 5-point Likert-type scale incorporating a 'don't know' response	64% response rate (335/525): item non-response was 5% on the WTP question compared with 18% on the VAS				Bid and health status significantly associated with WTP but none of the socio-economic variable were Possible range bias as upper bid set by authors to which 24% responded in the affirmative
552	Kartman *et al.*, 1996 Comparing CE with bidding game in the context of angina pectoral attacks	341/438 (78%) responses: 36 refused to participate, 41 protest answers and 20 missing values				Mean WTP for both methods was similar Evidence of internal validity, with mean WTP statistically significantly and positively related weekly attack rates and angina pectoris status Both methods indicate WTP increased with scope in the case of size of reduction in angina attacks. This was highly statistically significant in the bidding game but not statistically significant in CE method Starting point bias detected in the bidding game
553	Kartman *et al.*, 1996 Testing the CE WTP approach for scope and question order effects	400/461 patients responded (13% non-response rate) via telephone interviews by a nurse				WTP increased with the probability of being free from symptoms and reduced risk of relapse Method sensitive to changes in scope No question order effects were found
509	Kartman *et al.*, 1997 Testing the CE method with OE follow-up method for scope effects	Zero response for the OE question ranged from 7% to 25% and non-response ranged from 7% to 13%				Method insensitive to scope effects

continued

Willingness to pay (WTP) *contd*

Properties: has its theoretical basis in welfare economic theory[505] and is held to measure strength of preference.

Ref.	Empirical/methodological study	Acceptability to respondents	Cost	Internal consistency	Reproducibility	Validity
547	Kobelt, 1997 WTP for a reduction in incontinence symptoms CE method	85% response rate (excluding those who had difficulties in answering the WTP question led to reduction in usable responses to 66%)				Respondents willing to pay almost twice as much for double the improvement in symptoms. Proportion of patients willing to pay also increased for the higher level of benefit WTP positively and significantly associated with severity of symptoms
417	Krabbe *et al*., 1997 Comparability and reliability of SG, TTO, WTP and rating scales OE method			Using dominant pairs, incon-sistencies for WTP varied from 1.9% to 50.5% depending on distance between rates For rating scales, range = 0–12.5, SG = 0–21.6, TTO = 1.9–17.8 WTP Cronbach's α = 0.77 Rating Cronbach's α = 0.58 SG Cronbach's α = 0.65 TTO Cronbach's α = 0.49		Good agreement between ranking of health states and valuations: rating = 0.83, SG = 0.75, TTO = 0.77, WTP = 0.80 Good convergent validity between SG and TTO, but not so good for other comparisions
522	Lee *et al*., 1997 Patients' WTP for autologous blood donation CE method	Respondents entered into lottery to win $50 gift token. 44% response rate (235/528): 5.5% did not answer the WTP question, 94% rated the questionnaire 'easy or very easy'				WTP increased significantly with income, dread of allogenic transfusion, and perceived risk of requiring a blood transfusion

continued

Willingness to pay (WTP) *contd*

Properties: has its theoretical basis in welfare economic theory[505] and is held to measure strength of preference.

Ref.	Empirical/methodological study	Acceptability to respondents	Cost	Internal consistency	Reproducibility	Validity
518	Lindholm *et al*., 1997 WTP for prevention of cardiovascular disease Bidding game	Response rates 96% (100/104) for WTP question with 5% reduction in risk, and 94% (98/104) for the 10% reduction. 3 outliers treated as protests based on interviewers notes				Evidence concerning internal validity mixed, with some variables moving in the expected direction, and other moving opposite to *a priori* hypotheses
527	Loehman *et al*., 1979 Costs and benefits of air quality control in an urban region in Florida using WTP to value health effects PC approach	22.4% (404/1800) response rate				Tested on a separate sample of 47 college students and re-administered after 3 weeks found average correlation of 0.85 (range 0.82–0.95)
368	Miedzybrodzka *et al*., 1995 Women's preferences for cystic fibrosis screening method OE method	Of 450 responses, 173 (38%) gave inconsistent responses (i.e. were willing to pay less for their preferred option)				WTP for the stepwise method was significantly associated with social class. Being single (not married or living with someone) was significantly associated with WTP for the couple method; social class was not. Education was significantly associated with neither
631	Mills *et al.,* 1994 Fund raising for village activities in The Gambia					Stated WTP lower than cost of impregnating nets except in 1 region. Respondents likely to understate WTP in the hope of subsidies
548	Muller & Reutzel, 1984 WTP for a reduction in fatality risk through car crash protection	89% response rate (77/87)				Evidence found of scope effects. Respondents' WTP for twice the risk reduction was much greater A comparison of medians suggested a higher WTP when the risk was expressed out of 10,000 rather than 100, although the risk was the same Respondents' WTP was about 1/3 higher when asked to pay a monthly amount rather than an annual sum

continued

Willingness to pay (WTP) *contd*

Properties: has its theoretical basis in welfare economic theory[505] and is held to measure strength of preference.

Ref.	Empirical/methodological study	Acceptability to respondents	Cost	Internal consistency	Reproducibility	Validity
419	O'Brien & Viramontes, 1994 Comparing WTP (bidding game) with SG, short-form 36 (SF-36) and a rating scale	77% response rate (102/133 interviews: 16 could not be contacted, 13 refused to participate and 3 had language or hearing problems)			4-week test–retest reliability deemed acceptable at 0.66 (lower than SG = 0.82)	WTP significantly associated with household income No evidence of starting point bias SG was most highly related to WTP (r = 0.46) compared with a rating scale and SF-36
534	O'Brien *et al.*, 1998 Feasibility and validity of WTP for new drug for febrile neutropenia Bidding game approach, 2 algorithms	501 people replied they would be interested out of 3486 letters sent out. 220 interviewed. $25 incentive payment				WTP significantly associated with the size of risk reduction, although at a declining marginal rate. WTP higher in the lower prior risk group WTA exceeded WTP. 2 algorithms with the bidding game were used to test for bias. Although the second scenario gave a higher mean response, it was not statistically significant No starting point bias detected
262	Olsen & Donaldson, 1998 Using WTP to set priorities between 3 alternative programmes PC method	Subjects interviewed in their own homes by trained interviewers. 64% response rate (150/235): 7 excluded as would neither express WTP nor prioritise alternative programmes		Discrepancies between WTP and ordinal ranking. For 3 alternative programmes, although only 25% (35) gave strong ordering between all 3 in terms of WTP, 76% (109) did in terms of their ordinal ranking. Of the 40 respondents who were indifferent in terms of WTP across the 3 programmes, 27 gave strong preferences in terms of ordinal rankings		No evidence to support hypothesis that men would have higher WTP values for heart operations, although women had a higher WTP for hip operations. Education had a significant negative association with the helicopter ambulance programme but was not associated with either of the 2 other programmes Some evidence that length of interview time was positively associated with WTP WTP values for the helicopter programmes were similar to that charged for the actual programme in existence Only 25% of respondents had compared the different numbers of patients involved in each programme and for the other 75% it was the programmes themselves that were considered in the decision-making process[551]

continued

Willingness to pay (WTP) *contd*

Properties: has its theoretical basis in welfare economic theory[505] and is held to measure strength of preference.

Ref.	Empirical/methodological study	Acceptability to respondents	Cost	Internal consistency	Reproducibility	Validity
528	Pennie *et al.*, 1991 Empirical study incorporating students' WTP for a hepatitis B vaccine PC approach	95% of students who attended classes on the questionnaire day completed the questionnaire			Reliability was tested with a separate group of nursing students 2 weeks later and there were no significant differences	Students considering themselves at the least risk (as against higher-risk medical laboratory students) and therefore with the least to gain were the least willing to pay the retail price
353	Phillips *et al.*, 1997 WTP for poison control centres in the USA Bidding game approach	Response rates were 71% (of 396) people who had calls blocked due to budget cuts, 72% (of 418) who called after the block was lifted, and 48% (of 119) from the general population				Little evidence of free riding*, but some evidence of starting point bias
535	Ramsey *et al.*, 1997 WTP for anti-hyperi ntensive care PC method, although using a 5-point Likert scale instead of 'yes/no' for each bid (yes, definitely; yes, probably; no, probably; no, definitely; don't know)	Postal survey: 85% response rate (from 194) for payment arm (respondents were paid $3 each, although 10% of the money was returned) and 76% (of 206) from the non-payment arm. 30% gave an "uninterpretable" answer to the WTP question. Likelihood of uninterpretable answer increased with age. Design may have confused because some answered 'don't know' to every bid				Income and education increased as WTP increased. Lack of significant association between WTP and prior history of stroke, myocardial infarction and perceived health status
540	Reutzel & Furmaga, 1993 WTP for pharmacy services in a Veteran's hospital in the USA 2 payment vehicles used: participation costs (travel costs etc) and a fee	198 patients were interviewed. Mean age 63 years. 146 provided responses to both WTP questions				47/146 gave same WTP using both payment vehicles. The 2 methods did not give the same answers, but responses not statistically significantly different Whilst some evidence of internal validity, variance in WTP was unexplained. This may indicate that respondents 'guess' WTP values

* *Free riding refers to the tendency to underestimate or overestimate true WTP, depending on the incentives inherent in the question*

continued

Willingness to pay (WTP) *contd*

Properties: has its theoretical basis in welfare economic theory[505] and is held to measure strength of preference.

Ref.	Empirical/methodological study	Acceptability to respondents	Cost	Internal consistency	Reproducibility	Validity
260	Ryan *et al.*, 1997 Valuing alternative forms of antenatal care CE approach	704 of 956 (74%) returned the postal questionnaire. 658 (93%) of these completed the WTP section				Indirect evidence was found for the insensitivity of WTP to the type of care
559	Ryan *et al.,* 1998 Comparison of PC and CE approaches					CE WTP sensitive to treatment of 'don't know' responses but not non-demanders or level of integration
555	Ryan *et al.*, 1999 Comparing CE and with PC methods					The CE approach gave rise to significantly higher WTP estimates; may be due to 'yea-saying'
560	Ryan & Ratcliffe, 2000 Issues raised in analysing CE WTP data in context of antenatal care	93% (658/704) of respondents completed the WTP questions				WTP sensitive to the limits of integration, bid vector design and method of analysis
369	Ryan & San Miguel, 2000 Consistency of WTP estimates in context of 2 alternative surgical procedures for menorrhagia PC method	51% response rate (75/146)		41% (20) of patients preferred conservative surgery but were prepared to pay more for hysterectomy. 60% of respondents who gave inconsistent answers mentioned cost as the reason for their WTP		

continued

Willingness to pay (WTP) *contd*

Properties: has its theoretical basis in welfare economic theory[505] and is held to measure strength of preference.

Ref.	Empirical/methodological study	Acceptability to respondents	Cost	Internal consistency	Reproducibility	Validity
496	Swan *et al.*, 1999 Patients' utility from magnetic resonance and conventional angiography WTP, TTO, rating scales and other estimates collected. OE (% of income) approach	29 of the 30 respondents answered the WTP questions. 28/30 answered the TTO questions				There were no statistically significant differences between rating scale, WTP and TTO findings
508	Thompson *et al.*, 1984 Empirical study on feasibility of WTP for chronic arthritis OE approach	Response rate 27% (49 of 184), although respondents not pressed for answer. Respondents with more education, in paid employment or who were having more treatment for arthritis more likely to respond (rose to 71% for those in paid employment)			Test–retest reliability poor: correlation coefficient = 0.25, statistical significance = 0.08	WTP was positively associated with number of symptoms. WTP was "relatively" insensitive to income
544	Walraven, 1996 Empirical study of WTP for district hospital services in Tanzania OE approach	500 outpatients, 321 inpatients and 1500 householders were interviewed. In addition, 22 focus groups were conducted				A majority of patients actually paid more than they said they were willing to pay. 62% at 1 hospital and 67% at another. At another hospital, which later introduced user charges, WTP predicted behaviour "reasonably well"
632	Zeidner & Shechter, 1994 Students' WTP to reduce their level of anxiety during exams					Students with "higher stakes" (i.e. more to lose by being in a class with historically a higher fail rate) were more anxious and WTP more compared to students with "low stakes". Students who are more anxious during tests are WTP more to reduce their anxiety Students' WTP was associated with dissatisfaction with current anxiety levels, and students were WTP significantly more for larger than smaller reductions in anxiety Students' WTP was significantly less than they recommended others should pay

continued

Willingness to pay (WTP) *contd*

Properties: has its theoretical basis in welfare economic theory[505] and is held to measure strength of preference.

Ref.	Empirical/methodological study	Acceptability to respondents	Cost	Internal consistency	Reproducibility	Validity
495	Zethraeus *et al*., 1997 Women's quality of life with hormone replacement therapy for mild and severe menopausal symptoms Also investigates WTP for hormone replacement therapy treatment via CE method					May be range bias in analysis since cut-off upper value was used. Women with more severe symptoms had a higher WTP. WTP values were consistent with rating scale and TTO measurements

Measure of value (MoV)

Properties: through its nature of assigning numerical scores to alternative options, the technique of MoV incorporates a relative strength of preference measure. A constrained choice can also be applied if the selection of alternatives are priced, and respondents must make their choices within a given budget.[265,266]

Ref.	Empirical/methodological study	Acceptability to respondents	Cost	Internal consistency	Reproducibility	Validity
264	Churchman & Ackoff, 1954 The pioneering paper developing the technique and illustrating its potential usage	The nature of the technique may be quite a lengthy and tiresome process if there are many alternative options to compare				Validity should be encouraged by nature of the technique. Once preferences are ranked, they are examined again through a comparison against groups of alternatives and adjusted if necessary
265	Dickinson, 1979 Methodological paper on behalf of Hereford Health District to determine the best combination of priorities from a list of 11 'priced' healthcare goods within a fixed budget	Although agreeing the initial list of priorities was difficult, the author noted the time and effort involved in management time was around 15 hours, and was judged a "good buy". The author further noted that someone with a good knowledge and experience of the technique would be required as administrator				

Allocation of points

Properties: because the "budget" of points (tokens, money, lives) to allocate is fixed, respondents are subjected to a constrained choice.[276,277,561] Given the wording of the question, it has been argued to possess cardinal properties.[277]

Hoinville PEM*

Ref.	Empirical/methodological study	Acceptability to respondents	Cost	Internal consistency	Reproducibility	Validity
268	Hoinville, 1996 Methodological paper illustrating the PEM with examples				In the context of variables surrounding journeys to work, the retest 2 weeks later was 75% (of 120)	When price increased, a lower quantity was chosen, and *vice versa*
267	Hoinville, 1977 Methodological paper illustrating the use of the PEM with examples	During early development of the technique, only 3% of respondents had difficulty. No noticeable differences across socio-economic and age groups. Most enjoyed and got satisfaction from interview			32 respondents asked to repeat an exercise (using verbal and pictorial scales) after a 2-week interval. 24 gave consistent answers with their pictorial choice and 22 with their verbal choice. Using verbal/pictorial scale, out of another 32 respondents, interviewed a few weeks apart, 22 gave consistent answers. Finally, of 17 other respondents indicating their preferences on 5 3-point scales, 7 gave identical responses and 6 gave only 1 slight difference	16 respondents asked to allocate points between 5 3-point scales, on 2 occasions, 2–3 weeks apart. On the second occasion, the price of 1 scale was increased and another decreased. 4 respondents gave the same answers whilst 9 reacted positively to the price change. From a further 611 respondents in a computer study only 6% acted irrationally by buying more when the price went up and less when it went down. In another study, of 121 respondents, only 1% made illogical choices When a group of respondents were told they had to act on behalf of local councils compared to those acting for themselves, only 1 out of 10 variables was significantly different at the 5% level

* *The Hoinville PEM is one of the first applications of the allocation of points approach*

Allocation of points *contd*

Properties: because the "budget" of points (tokens, money, lives) to allocate is fixed, respondents are subjected to a constrained choice.[276,277,561] Given the wording of the question, it has been argued to possess cardinal properties.[277]

Allocation of points

Ref.	Empirical/methodological study	Acceptability to respondents	Cost	Internal consistency	Reproducibility	Validity
633	Bate, 1999 Importance of public views against other criteria for priority setting Comparison of choice-based CA and budget pie (allocation of points)	Response rate to a postal questionnaire without reminders was 29.4% (42/143). 11 were unsuccessfully completed leaving 31 usable responses (21.7%)				Variation in relative importance of criterion across methods. However, lack of numbers suggest further work
562	Srivastava *et al*., 1995 A value hierarchy (similar to the budget pie where 100 points were allocated between attributes) and ranking systems				Reliability = 0.64–0.69, but small numbers	High correlation between swing weights, ranking and rank order centroid
277	Clark, 1974 Methodological paper on the budget pie	Reports unpublished work by Ostrom, which states middle and upper status respondents have little difficulty using this approach. After trained interviewers explained, response rates of 80–90% have been achieved with lower status respondents				
282	Strauss & Hughes, 1976 Tax expenditure recommendations by residents of North Carolina. Budget pie method with allocation of coupons and budget constraint	Response rate to a postal questionnaire was 28.5% (1001/3517) after 1 reminder and a 5-week cut-off. Method described as "simple and inexpensive"	Small cost involved (£6000) for self-administered questionnaire			
281	Ratcliffe, 1999 Allocation of donor liver grafts	Response rate of 38% to a convenience sample 14% thought difficult, 27% moderately difficult and 22% slightly difficult				Internal validity confirmed Of 303 respondents, 2 exhibited dominant preferences by consistently allocating all 100 livers to the group of individuals with the highest expected length of survival

Allocation of points *contd*

Properties: because the "budget" of points (tokens, money, lives) to allocate is fixed, respondents are subjected to a constrained choice.[276,277,561] Given the wording of the question, it has been argued to possess cardinal properties.[277]

Schedule for the evaluation of individual quality of life – direct weighting (SEIQoL–DW)

Ref.	Empirical/methodological study	Acceptability to respondents	Cost	Internal consistency	Reproducibility	Validity
274	Hickey *et al.*, 1996 Empirical application of SEIQoL–DW to HIV/AIDS	Quick and practical in clinic				
275	Browne *et al.*, 1997 Development of the SEIQoL–DW procedure (and a comparison with SEIQoL)	More reliable than SEIQoL. DW took 5 minutes to understand and complete compared with 10–30 minutes for SEIQoL (and slightly less time 7–10 days later)		r = 0.73 (SEIQoL)	Weights derived by SEIQoL–DW changed on average by 4.5 points compared to 8.4 with SEIQoL (7–10 days later) κ = 0.51 (SEIQoL–DW) κ = 0.31 (SEIQoL) 14–20 days later 55% of respondents could identify the weights they provided in SEIQoL–DW	Weights produced by SEIQoL and SEIQoL–DW differed: at time 1 the mean difference was 7.8 points and at time 2 7.2 points R^2 = 0.78 (SEIQoL)
565	Browne *et al.*, 1997 Conceptual overview of quality of life assessment including *inter alia* SEIQoL	SEIQoL–DW was adapted for use (from SEIQoL) when rapid answers required				

Allocation of points *contd*

Properties: because the "budget" of points (tokens, money, lives) to allocate is fixed, respondents are subjected to a constrained choice.[276,277,561] Given the wording of the question, it has been argued to possess cardinal properties.[277]

Patient-generated index (PGI)

Ref.	Empirical/methodological study	Acceptability to respondents	Cost	Internal consistency	Reproducibility	Validity
269	Ruta *et al.*, 1999 Postal PGI for measuring quality of life. Combines patient self-assessment and weighting	74% response rate (571/777). Of these, 63% completion (359). Pilot: 47% fully and correctly complete			Test–retest coefficient = 0.7 at 2 weeks (more reliable than SF-36 for group comparisons)	High correlation with a clinical (low back pain) measure and SF-36 Referred patients had significantly lower (worse) score than those managed in general practice
270	Macduff & Russell, 1998 Use of PGI and SF-36 to measure patient quality if life in a disabled population over time	81% response rate (131/161): 62% usable rate at time 1, 52% at time 2			Good test–retest but low sample eligible for retesting because of inaccurate completion	Box 6 is intended to relate problems to rest of life as basis for weighting. Results suggest that it is not understood by respondents and therefore the PGI is not assessing overall quality of life
271	Ruta *et al.,* 1999 Methodological paper evaluating the reliability, validity and responsiveness of the PGI in 4 conditions	75.4% response rate (1317/1746); 51% (672) of these correctly completed			Test–retest coefficient = 0.65 at 2 weeks	PGI correlated with 4 condition-specific scores
272	Herd *et al.*, 1997 Measurement of quality of life in atopic dermatitis patients using PGI and Dermatology Life Quality Index					Correlated 'moderately well' with Dermatology Life Quality Index and PGI correctly stressed the worst problems. PGI elicited problems not identified by structured Dermatology Life Quality Index, and scored quality of life worse than did Dermatology Life Quality Index
273	Jenkinson *et al.*, 1998 Comparison of 3 techniques for measuring patient (self-reported) health status: PGI, EuroQoL and SF-36	89/100 completion with PGI and EuroQoL; 86/100 with SF-36				EuroQoL not correlated with SF-36 or PGI PGI by far the most sensitive to change over time

Appendix 3

Additional data for chapter 6

We have provided a summary of some of the studies identified in the review to provide the reader with a summary of the main issues raised for each of the qualitative techniques. Note: this table does not provide a listing of all the studies identified in our review.

Interviews

Reference	Study	Description	Methodological comments
291	Crabtree & Miller, 1991 A qualitative approach to primary care research: the long interview	4 patients with recent experience of pain and 4 physicians were interviewed about pain perception. This paper details the research process including questionnaire development, transcription, identifying utterances and identifying emerging themes	A purposive sample was selected to ensure that respondents and interviewers were strangers and that respondents did not have a specialised knowledge of the topic. The time scale was considered to be acceptable and the cost was considered in relation to the time needed to complete the study to a high standard. Each researcher wrote down their experiences of pain in order to understand biases that may be introduced rather than to eliminate bias. Also, 3 researchers were involved with independently reading interview scripts. The study is stated as being generalisable and transferable
287	Foster, 1994 Fishing with the net for research data	A number of teachers and lecturers were approached via email to establish how they go about planning and designing their courses	Advantages include saving time and money because of the elimination of transcription of interview tapes, and no need for travel to participants' homes or offices. Disadvantages are that misunderstandings cannot be instantly clarified, and interesting emerging issues cannot be probed. Also, mail may be viewed as being a nuisance to busy professionals
289	Dicker & Armstrong, 1995 Patients' views of priority setting in health care: an interview survey in one practice	16 patients from one general practice were interviewed for 30–40 minutes, exploring underlying assumptions of attitudes towards rationing in the NHS	The sample was chosen to provide a cross-section of the types of people attending the practice, rather than seeking to be representative. One person identified themes from the transcripts. Patients found it unacceptable to discuss issues in relation to themselves and found it easier to consider other people's needs, they also felt ill equipped to answer priority-setting questions in general
290	Williams *et al.*, 1998 The meaning of patient satisfaction: an explanation of high reported levels	15 people referred to a community mental health team were interviewed both before and after their initial consultations to establish their satisfaction with the service they received	A small sample of people participated (no explanation of sample size or recruitment is provided). Attention is paid to the location of the interviews: it was thought that conducting them in the homes of the patients was advantageous. Second interviews were undertaken so as to provide an opportunity for respondent validation (no mention of who or how many participated in the analysis stage)
292	Wilson *et al.*, 1998 Telephone or face-to-face interviews?: a decision made on the basis of a pilot study	20 interviews were undertaken as the pilot study to ascertain if telephone or face-to-face interviews would be appropriate for examining continence care. De Vaus's* 5 considerations were used to aid in the decision-making process	Telephone interviews were perceived to be the most appropriate method for the main study. Advantages were response rate, cost-effectiveness, less time needed for each interview, quality of responses and interviewer safety
288	Ayanian *et al.*, 1999 The effect of patients' preferences on racial differences in access to renal transplantation	1392 patients receiving care for end stage renal disease were interviewed about preferences concerning transplantation	Sample may not be representative

* *De Vaus DA. Surveys in social research. 3rd ed. London: University College London Press; 1991 (cited in Wilson and colleagues[292])*

Delphi technique

Reference	Study	Description	Methodological comments
302	Thomson & Ponder, 1979 Use of Delphi methodology to generate a survey instrument to identify priorities for state allied health associations	Used a 3-stage Delphi to develop a survey technique for use in identifying priorities for the Texas Society of Allied Health Professions (TSAHP). 21 health professionals were recruited and were mainly health educators or administrators	The study was piloted to minimise ambiguities, but selection of participants was open to biases. Data to be eliminated was reported and reasons for this provided. Stated as being time-consuming, and therefore costly plus may be unacceptable to researchers in some situations
297	Charlton *et al.*, 1981 Spending priorities in Kent: a Delphi study	Delphi technique used by the Kent Area Health Authority to elicit spending priorities of all those involved in health service decision-making (professional and lay policy-makers) in Kent	Ascertains that different professionals adhere to their initial responses regardless of being aware of the opinions of other groups, i.e. anonymity increases validity? Cost is stated as being comparatively low
300	Gabbay & Francis, 1988 How much day surgery? Delphic predictions	2 panels (1 national consisting of 9 panelists, 1 local consisting of 19 panelists) were asked to estimate the probable rates of day surgery for 83 procedures, using Delphi. These results were then compared with each other and with a Hospital Activity Analysis	The study is evaluated in the discussion, and limitations pointed out. It highlights problems with generalisability of the national Delphi study. Surgeons did not comment on subspecialist operations, therefore the results for these procedures may be less reliable
304	Burns *et al.*, 1990 Primary health care needs of persons with physical disabilities: what are the research and service priorities?	3-stage Delphi asked a variety of health professionals and members of disability organisations to establish a consensus on disability service and research priorities	Not clear how many disability organisation members were included in the sample. Reliability mentioned for testing correlation of results for rounds 2 and 3, but not mentioned in the context of predefined criteria
634	Silva & Lewis, 1991 Ethics, policy, and allocation of scarce resources in nursing service adminis-tration: a pilot study	A 3-round Delphi technique was used to ask 31 student nurses to identify and prioritise issues of ethical concern that stem from the allocation of scarce resources	
296	Harrington, 1993 Research priorities in occupational medicine: a survey of United Kingdom medical opinion by the Delphi technique	Delphi technique applied to gain opinions of 53 senior practitioners concerning occupational medicine research priorities	Sample limited by targeting a very specific group; improvements could be made by asking multidisciplinary team. Consensus reached but when checked by respondent validation some results were found to be untrustworthy

continued

Delphi technique *contd*

Reference	Study	Description	Methodological comments
305	Kastein *et al.*, 1993 Delphi, the issue of reliability: a qualitative Delphi study in primary health care in the Netherlands	The primary aim of this study was to develop a set of criteria for evaluating the performance of family physicians consulted about abdominal pain. Additionally, the secondary objective was to explore issues of reliability in the Delphi technique	Explores issues of situation and person-specific biases, and makes and explains methods of doing this for their study. Does not deal with all types of reliability and omits to examine other criteria such as validity, acceptability and objectivity
298	Roberts *et al.*, 1994 Setting priorities for measures of performance for geriatric medical services	89 consultant geriatricians were asked to suggest measures appropriate for performance indicators in a 2-round Delphi study and 44 patients were interviewed and asked to rank these in order of importance and suggest other measures. The rank order of the 2 groups is compared	No internal validation. Patients were interviewed only once; no follow-up due to practicalities of locating the patients
303	Gallagher *et al.*, 1996 Policy priorities in diabetes care: a Delphi study	28 experts in diabetes, including patients and patient representatives were asked to set priorities for improving care for diabetic patients, using a 2-round Delphi survey	Researcher bias minimised ensuring objectivity; inclusion criteria for participants formulated; attempts were made to include non-responders. General considerations made about technique such as consensus methods tend to ignore dissenting contributions and that poor selection of participants can result in poor reliability
299	Hadorn & Holmes, 1997 The New Zealand priority criteria project. Part 1: Overview	2-stage Delphi process asking different types of health professionals to set criteria for elective surgical procedures	Inclusion criteria and participating numbers are not made clear; this is true of patients and members of the public as well as for professional groups. It is therefore difficult to establish the validity, reliability and generalisability of this application of the technique. It should be stressed, however, that this is an overview
301	Wilson & Kerr, 1998 An exploration of Canadian social values relevant to health care	4 Delphi-style surveys were conducted to gain opinions from 209 members of a bioethics society and 144 persons believed to be knowledgeable about the Canadian healthcare system about social values in relation to healthcare	The sample obtained is stated as not being representative: it consisted of participants who were well educated and affluent. This limits the study in terms of generalisability. Also, not all respondents replied to all 4 stages, although numbers are not made clear
635	Campbell *et al.*, 1999 The effect of panel membership and feedback on ratings in a 2-round Delphi survey: results of a randomized controlled trial	A 2-round postal Delphi survey involving both healthcare managers and family practitioners on issues of quality of primary healthcare	Addresses similar issues as above. Concludes that panel composition can influence the types of decision made, questioning the reliability of the method
295	Endacott *et al.*, 1999 Can the needs of the critically ill child be identified using scenarios? Experiences of a modified Delphi study	The Delphi technique is used to gain expert opinion (paediatric intensive care sisters) of the needs of children admitted to the intensive care unit. Scenarios are presented to the panel and are asked to indicate the importance of each need. The technique itself is evaluated	Reports the method as having strong validity and no evidence of reliability. In this application both face and content validity are established. Demanded a high workload, and therefore is not acceptable to all those approached, which lowered the response rate

Focus groups

Reference	Study	Description	Methodological comments
320	Keller *et al.*, 1987 Assessing beliefs about and needs of senior citizens using the focus group interview: a qualitative approach	22 individuals aged 65 and over, and 16 people aged between 22 and 40 were asked to express their views on the healthcare needs of older people. The second group was selected because they were perceived to be knowledgeable about the structure and dynamics of the communities concerned	A 'non-probability' sample was used. Telephone follow-ups were carried out to give the opportunity for participant feedback to validate what they had said during the groups. It is unclear who undertook the analysis stage and if more than one researcher was involved
314	Kitzinger, 1994 The methodology of focus groups: the importance of interaction between research participants	351 people participated in one of 52 focus groups. Looked at different groups' perspectives on AIDS	Groups were pre-existing to establish how people talk about AIDS with peers, i.e. how they might naturally discuss this issue. Everyone was encouraged to talk. This was achieved by playing a game at the beginning of each session
312	Kuder & Roeder, 1995 Attitudes toward age-based health care rationing: a qualitative assessment	46 individuals (19 male and 27 female) took part in focus group discussions. They were grouped by age and socio-economic status. The groups were presented with a series of scenarios and asked how they would distribute scarce resources	Gender and race were not considered when selecting group members. Analysis and coding were carried out by more than one member of staff
317	Powell *et al.*, 1996 Focus groups in mental health research: enhancing the validity of user and provider questionnaires	Asks both users and providers of mental health services in 4 separate focus groups to generate ideas for what should be included in questionnaires and to validate what is in existing questionnaires	The user groups were self-selecting, and therefore important views of less confident people may have been missed. Also, they were not stratified for variables such as age, sex and occupation. The provider groups consisted of staff of equivalent levels so as not to be intimidating, but also may have missed important information. 2 researchers were involved in the analysis stage. States the study has limited generalisability. The authors discuss all limitations of the study
319	Stevens, 1996 Focus groups: collecting aggregate level data to understand community health phenomena	13 lesbian women were asked about their healthcare experiences and their interactions with health professionals	Free flowing discussion occurred, enabling topics to be dealt with in depth. Only those who were very well educated were included and therefore yielding a bias sample. Analysis carried out by the facilitator of the groups
325	Weinberger *et al.,* 1998 Can raters consistently evaluate the content of focus groups?	101 heart disease patients and 29 heart physicians were asked about their experience of healthcare and factors affecting decision-making, respectively. 3 raters were asked to categorise lists formulated from the transcripts so as to establish the level of agreement between them	It was found that consistency between the raters was difficult to achieve

continued

Focus groups *contd*

Reference	Study	Description	Methodological comments
322	Bradley *et al.*, 1999 The health of their nation: how would citizens develop England's health strategy?	Conducted in 4 general practices and 1 school. Included 173 people in 24 focus groups and asked about citizens' attitudes towards England's health strategy	Analysis stage carried out by 2 independent qualitative researchers. Cost stated as being 'inexpensive' at £365 per 90-minute focus group. Participants were able to grasp themes and found it acceptable to discuss topics put forward
313	Cohen & Garrett, 1999 Breaking the rules: a group work perspective on focus group research	Asked groups of people with mental illness about patient/worker relationships. Uses examples of focus groups to demonstrate that sometimes sensitive issues can be explored	Again, used pre-existing groups who were used to discussing their situations. The facilitator was able to establish a trusting relationship with the participants, which enabled group members to feel comfortable and talk openly
323	Dolan *et al.*, 1999 Effect of discussion and deliberation on the public's views of priority setting in health care: focus group study	60 people were selected using stratified random sampling. The aim was to establish how much attitudes and opinions changed after a period of discussion and deliberation was allowed	Accurate and considered opinion obtained due to the period of deliberation provided. The same 2 researchers conducted all meetings to ensure consistency and that all relevant topics were covered adequately. The meetings were all tape-recorded and transcribed, although it was unclear who carried this out

Citizens' juries

Reference	Study	Description	Methodological comments
334	Dunkerley & Glasner, 1998 Empowering the public? Citizens' juries and the new genetic technologies	Describes a citizens' jury that took place in Wales, which looked at new technologies available for genetic testing for common disorders	Both organisers and jurors commented that an unrepresentative sample of people had been selected. All but one had left school at the minimum leaving age, few were in full-time employment, there was no representative from ethnic minority groups; there was also no spread of age. The objectivity of the moderator cannot be established but is stated as having experience in facilitating small group discussion. It was felt that some of the jurors did not fully understand all of the material that was being discussed

Consensus panels

Reference	Study	Description	Methodological comments
336	Stronks *et al*., 1997 Who should decide? Qualitative analysis of panel data from public, patients, healthcare professionals, and insurers on priorities in health care	5 panels were chosen: each one selected to represent a distinct group. These groups were patients, the public, GPs, specialists and health insurers. Panel members were asked to prioritise 10 different health services	The panel representing the public was selected from university students and civil servants not working in health. This cannot be said to be representative of the public. The authors also state that none of these had a chronic illness or handicap. Authors also state that those people taking part had neither adequate knowledge of the subject matter nor time to deliberate these matters

Public meetings

Reference	Study	Description	Methodological comments
341	Gundry & Heberlein, 1984 Do public meetings represent the public?	Uses examples of 3 public meetings and simultaneous surveys to test 3 different hypotheses relating to the likely representativeness of those attending the meetings	Although representativeness of public meetings cannot be established, opinion gained at the meetings generally matched opinion gained through conducting wider surveys
343	Gott & Warren, 1991 Neighbourhood health forums: local democracy at work	Several public meetings were set up and reviewed	Participation encouraged by personally inviting members of the community, having a diverse range of professionals attending and by holding the meeting in a well-known building

Appendix 4

Establishing the relative weights of methodological criteria when evaluating research techniques

Background

Current systematic reviews of methodologies have taken a qualitative approach, highlighting how instruments perform against predefined criteria. This study represents an attempt to estimate the weights of such predefined criteria.

Methods

Following the identification of criteria for assessing the methodological status of techniques, as defined in chapter 4, a choice-based CA exercise was conducted to establish both the weights of these criteria and also whether differences existed across disciplines. The criteria included in the study, and their corresponding levels, are shown in *Table 8*.

The combinations of criteria and levels resulted in 2916 ($2^2 \times 3^6$) possible definitions of instruments. The computer software package SPEED[636] was used to reduce this to 28. These 28 scenarios were randomly paired into 14 choices. For each choice the respondent had to choose between instrument A and B (*Figure 1*). Two scenarios were presented: one dealing with an allocative efficiency question (eliciting public view concerning which one of three different treatments should be allocated scarce funds), the other with a technical efficiency question (eliciting the public's view of alternative ways of providing an asthma clinic). Two of the 14 choices (one in each scenario) were selected to gauge the internal consistency of the responses. In these two choices, all the criteria levels for one instrument were 'better', meaning that the respondent should choose that option. Respondents who failed both tests of consistency were assumed to be answering 'irrationally', or not taking the questionnaire seriously, and were dropped from the analysis.

The choice CA postal questionnaire (after successive piloting and modification) was sent to five distinct disciplines: HEs, health service researchers (HSRs), health council members (HCMs), medical sociologists (MSs) and PHCs.

From the responses to the 14 choices for each respondent, the following equation was estimated:

$$\Delta V = \alpha_1 ACCEPT + \alpha_2 CHOICE + \alpha_3 COST + \alpha_4 INTERNAL + \alpha_5 REPROD + \alpha_6 STRENGTH + \alpha_7 THEORY + \alpha_8 VALIDITY + e$$

where ΔV is the change in utility (or benefit) of moving from instrument A to B, and the independent variables are the differences in the levels of the criteria of the two instruments, as defined in *Table 8*. The α_1 to α_8 represent the weights of the criteria. These weights indicate

TABLE 8 *Criteria and levels for the discrete choice CA study*

Criteria	Levels	Coding
Acceptability to respondents – ACCEPT	Low	1
	Medium	2
	High	3
Whether a constrained choice is offered – CHOICE	No	0
	Yes	1
Cost – COST	Low	1
	Medium	2
	High	3
Internal consistency – INTERNAL	Low	1
	Medium	2
	High	3
Reproducibility – REPROD	Low	1
	Medium	2
	High	3
Strength of preference – STRENGTH	No	0
	Yes	1
Theoretical basis – THEORY	No	0
	Yes – from other discipline	1
	Yes – from own discipline	2
Validity – VALIDITY	Low	1
	Medium	2
	High	3

Choice 1	Instrument A	Instrument B
Acceptability to respondents	High	Medium
Constrained choice	Yes	Yes
Cost	Medium	Low
Internal consistency	High	Low
Reproducibility	High	Low
Strength of preference	No	Yes
Theoretical basis	Other discipline	Own discipline
Validity	Medium	Low
Which option would you prefer? (*tick one box only*)	Prefer instrument A ☐	Prefer instrument B ☐

FIGURE 1 *Example of a choice set in the weighting exercise*

the marginal change in overall utility, *V*, resulting from a marginal change in a given criterion. This marginal change is obviously dependent on the unit of measurement. So, for example, whilst the coefficient on acceptability to respondents indicates the marginal change in benefit of moving from say 'medium' to 'high', the coefficient on 'constrained choice' indicates the marginal change in benefit of moving from 'no' to 'yes'. The *e* represents the unobservable error term in the model. Given that individuals provide multiple observations, a random effects probit model was used to analyse the responses. A general to specific approach was adopted, with criteria excluded in a backward stepwise fashion if they were not statistically significant at the 5% level. Analysis was carried out first on the full data set and then by discipline. For the segmented analysis, the Chow-type Likelihood Ratio test was used to investigate whether the weights for the different criteria differed according to the scenarios presented. If there was no difference the data sets could be merged and jointly analysed according to discipline.

Results

From a total of 1227 questionnaires sent out, 690 were returned in the required timescale (which included one reminder). Of these, 144 were returned uncompleted and seven gave internally inconsistent responses. The remaining sample frame consisted of 539 questionnaires, giving a response rate of 56%. *Table 9* provides the background characteristics of respondents.

TABLE 9 *Sample frame on which the analysis was based*

Discipline	Frequency	Percentage (%)
Health economics	156	28.9
Health service research	61	11.3
Health council members	111	20.6
Medical sociology	99	18.4
Public health	112	20.8

For the total sample all the criteria had the expected sign and were significant at the 5% level (*Table 10*). This suggests that all the criteria are important when choosing between techniques. The positive values of seven of the eight criteria indicate that the higher these are in instrument B relative to A, the more likely the respondent is to choose instrument B. Similarly, the negative value of 'cost' indicates that the lower the cost, the more likely the individual is to choose the instrument. These results are all what we would expect, providing support for the internal validity of the discrete choice exercise. The weight of each criterion is indicated by the coefficient. In *Table 10,* the coefficients are reported in order of their relative importance. As mentioned above, it is important to consider the unit of measurement when interpreting weights. For the combined group a marginal change in validity is more important than a marginal change in any other criteria. This is followed by acceptability to respondents and strength of preference. The least important criteria are internal consistency and theoretical basis.

TABLE 10 *Weights for the criteria*

Criteria	Coefficient (weight)	*p*-value
Validity	0.4854	0.001
Acceptability to respondents	0.4790	0.001
Strength of preference	0.2348	0.001
Reproducibility	0.2231	0.001
Constrained choice	0.2149	0.001
Cost	–0.1769	0.001
Internal consistency	0.1354	0.001
Theoretical basis	0.0485	0.017
Number of observations	7346	
Log-likelihood function	–3166.5537	
AIC[a]	0.8643	

[a] *Refer to* Table 3 *footnote (page 60)*

The Chow-type likelihood ratio test did not reject the hypothesis of homogeneity (the weights are the same irrespective of the scenario) for all but the sample of HEs. The data for the two scenarios could therefore be merged for all groups except HEs. The specific results for the five disciplines are shown in *Table 11,* and *Table 12* shows the rankings derived from the weights. It is important to remember the unit of measurement when interpreting these weights.

Discussion and conclusion

The results from this weighting exercise suggest that the criteria identified for evaluating quantitative instruments (from the literature review in the main body of the report) reflect the criteria that methodologists see as important when choosing a technique. This finding supports the evaluation of the techniques identified (again, in the main body of the report) according to the predefined criteria.

The aggregated results show that all the criteria are important to methodologists when choosing a technique. Only HEs made a distinction between the scenarios. Given that HEs made up a large proportion of the sample frame (*Table 9*), this group may bias the overall results. Given this, analysis was carried out for the individual groups. When this was done, only one discipline (MSs) viewed all eight criteria as important. Acceptability to respondents and validity were constantly ranked highly by all disciplines, except by PHCs for the latter. Internal consistency and theoretical validity (which was rarely significant) were consistently ranked the lowest and constrained choice and cost fluctuated amongst the groups.

It is acknowledged that the robustness of these results may be affected by the relatively

TABLE 11 *Weights by discipline and scenario (for HEs only)*

Criteria	Coefficient (weight)					
	HE		HSR	HCM	MS	PHC
	Scenario 1	Scenario 2				
Acceptability to respondents	0.4799**	0.5979**	0.3777**	0.4899**	0.4911**	0.6722**
Constrained choice	0.6916**	0.0873*	0.1334*	0.1283**	0.5035**	0.7456**
Cost	–0.2280**	–0.2294**	ns	–0.2664**	–0.1607*	–0.5430**
Internal consistency	0.2364**	0.1942**	ns	0.1240**	0.1840**	0.2416**
Reproducibility	0.3770**	ns	0.3672**	0.2037**	0.2509**	0.3136**
Strength of preference	0.4624**	0.3734**	0.2993**	0.1307*	0.2076*	ns
Theoretical basis	ns	0.1170**	ns	ns	0.1232*	0.0897*
Validity	0.5017**	0.3954**	0.7959**	0.2528**	0.6412**	0.2317**
Number of observations	1074	1068	813	1524	1334	1519
Log-likelihood function	–414.9798	–478.8963	–324.4939	–737.1041	–509.0744	–587.2561
AIC[a]	0.7858	0.9099	0.8106	0.9765	0.7752	0.7824

[a] *Refer to* Table 3 *footnote (page 60)*

ns, not significant

** p < *0.01;* * p < *0.05*

TABLE 12 *Summary of the relative importance rankings obtained from the CA results*

Criteria	Aggregated	Segmented by discipline					
		HE		HSR	HCM	MS	PHC
		Scenario 1	Scenario 2				
Acceptability to respondents	2	3	1	2	1	3	2
Constrained choice	5	1	7	5	6	2	1
Cost	6	7	4	ns	2	7	3
Internal consistency	7	6	5	ns	7	6	5
Reproducibility	4	5	ns	3	4	4	4
Strength of preference	3	4	3	4	5	5	ns
Theoretical basis	8	ns	6	ns	ns	8	7
Validity	1	2	2	1	3	1	6

low response rate (56%), with possible biases introduced through non-response. There is also an issue about the complex task given the inclusion of eight attributes, the subjective interpretation of these 'qualitatively' defined attributes and the assumption of a linear additive model. Nevertheless, these findings are potentially useful for a number of reasons. First, they may help explain why certain disciplines choose certain research techniques. Secondly, the estimated weights could potentially be used to 'score' instruments in terms of their methodological status. Therefore, future research could develop this approach to carrying out a methodological systematic review, taking account of the above limitations of this study.

Appendix 5

Summary of criteria used in priority setting and priority-setting frameworks

Area	Criteria	Priority-setting framework	Involvement of public views?
England: City and Hackney[280,637]	• Robustness or the extent to which the proposal can be implemented (0–3) • Promotion of equity (0–1) • Evidence of effectiveness or cost-effectiveness (0–2) • Collaboration or integration with primary care (0–3) • Prioritised by the CHC (0–1) • Prioritised by local GPs (0–1) • Other possible or more appropriate sources of funding (0–5)	A 2-stage scoring system was used whereby first the proposals were ranked according to which responded greatest to local needs They were then scored further using the criteria (possible scores shown in brackets) Weights were determined by the director of public health in consultation with purchasing team managers and members of the health authority (not stated how weights were determined) Proposals ranked by multiplying the score by the weights	✔
England: Southampton[637]	• Health gain • Equity • Local access • Personal responsibility • Choice	Weights ascertained using an allocation of points exercise: 100 points to allocate across 5 criteria Proposals then scored according to the amount of health gain expected (score of 1–3) and whether or not other criteria were met (score of 0 or 1) Each of the proposals was then ranked by multiplying each criteria's weight by the associated score, and summing	?
Scotland: Argyll and Clyde[638]	• Potential health gain • Prevention of ill health • Quality of life • Equity of access • Addressing health status inequalities at population level • Expressed demand, appropriateness • Strength of evidence • Known priorities • Additional cost per person receiving the intervention	Scoring and weighting in a single stage. Each criterion is allocated a range of possible scores. For example, criterion A may be allocated a maximum score of 10 and a minimum score of –5, compared with criterion B which may have a range from 5 to –5. By definition, criterion A is more important than B Health proposals are then scored for each criterion according to this range, and a total score is estimated for each proposal	✗
Scotland: Ayrshire and Arran[639]	• Health gain • Effectiveness • Equity • Public preference • Flexibility • Value for money	Each of the criteria is first scored out of 10 and then multiplied by its relative importance (weighting), which are decided upon by executive and non-executive directors of the health board The weights are determined by an allocation-of-points method. Respondents are given 60 points to allocate across the 6 criteria Each of the proposals are then ranked by multiplying the score by the weight and summing	✔

continued

continued

Area	Criteria	Priority-setting framework	Involvement of public views?
Scotland: Greater Glasgow (Walker A, personal communication, 1999)	• Size of health gain • Quality of evidence to support change • Fit with local and national priorities • Enhance service quality (other than health gain) • Facilitate provision in primary care Recommended scores ranging from 0 to 10 were assigned to possible levels for each of the criteria	As above, each of the criteria are scored out of 10 and then multiplied by their weighting, which was agreed upon by a collaboration of health board directors and HEs The weights are determined by an allocation-of-points method. Respondents are given 100 points to allocate across the 5 criteria Each of the proposals are then ranked by multiplying the score by the weight and summing	✗
New Zealand[299,640–643]	Essential: • the health issue has a significant impact on the current and future health status of the population • the health issue promotes population-based methods to protect/ prevent health High weighting: • the health issue will reduce inequalities in health status • the health issue promotes the best health gain for the resources required Medium weighting: • there is public support for the health issue	Using these criteria, members of the Public Health Commission assessed the health proposals, although there is no mention of quantitative values assigned to each of the criteria	✔
Scotland: Aberdeen Royal Hospitals Trust[211]	• Level of evidence of clinical effectiveness • Size of health gain • Contribution to professional development • Contribution to education, training and research • Strategy area	Consultants working within the hospital trust were given a choice-based questionnaire, and the weights for the criteria were indirectly estimated Clinical directorates were then asked to score their proposed clinical service developments according to how well they performed on each of the criteria These scores were multiplied by the weights, and a total score estimated for each proposal	✗
Scotland: Dumfries and Galloway[644]	• Evidence of effectiveness • Value for money • Health gain or maintenance • Equity • Matching a national priority or board priority • Public preferences	Proposals are first scored according to each criterion on a 1–5 scale Each criterion is given a weight between 1 and 10 Each of the proposals are then ranked by multiplying the score by the weight and summing	✔
The Netherlands[637]	• Is the care necessary from the community point of view? • If so, has it been demonstrated to be effective? • Is it also efficient, using such methods as QALYs? • Can it still be left to individual responsibility?	Proposals had to pass through each of the criteria before they could be implemented	✔

continued

continued

Area	Criteria	Priority-setting framework	Involvement of public views?
Oregon[283–286,645–648]	Value to society: • prevention, many benefits, impact on society, quality of life, personal responsibility, cost-effectiveness, community compassion, mental health and chemical dependency Value to an individual needing the service: • prevention, quality of life, ability to function, length of life, mental health and chemical dependency, equity, effectiveness of treatment, personal choice community compassion Essential to basic healthcare: • prevention, many benefits, quality of life, cost-effectiveness, impact on society	Commissioners divided 100 points between the 3 main criteria: value to society, value to an individual needing the service and essential to a healthcare package In all, 17 proposals were then scored against these 3 criteria on a scale of 1–10 The weights are multiplied by the rating score to derive a total score for each proposal	✔ ?

Appendix 6

Questionnaire

Preferences for priority-setting criteria: developing a scoring and weighting system for health board/authorities

We are conducting a research project for the NHS R&D Health Technology Assessment programme which is aimed at identifying the relative importance of criteria that are commonly used to aid the priority-setting process. This study considers the potential for the approach employed in the paper to be used to establish weights for the criterion such that a score can be developed which encompasses these.

This questionnaire is concerned with your personal/independent views and the importance you place on the various criteria that are considered in this study. We would therefore be grateful if you could take the time to fill in this questionnaire. Everyone's responses and opinions are important. Once it is completed it should be returned in the prepaid envelope (enclosed) by 24 August 1999. If you do not wish to respond please return it uncompleted.

All answers and data will remain confidential and will not be reported in ways that can identify individual responses. If you have any queries about the study or the questionnaire, please contact Angela Bate at the Health Economics Research Unit, Department of Public Health, University of Aberdeen. Telephone 01224 663123, ext. 52783.

Thank you in advance for your assistance.

Angela Bate Mandy Ryan

Research Assistant MRC Senior Fellow

This questionnaire contains questions relating to the importance of the criteria in the priority-setting process. Each one presents the criteria (defined below) that are commonly used to analyse proposals and priorities. Care should therefore be taken to read the specific definition and context within which these criteria and their corresponding levels are set. Whilst some of the criteria levels may be crude, they will provide useful information on the relative weights of the different criteria.

The questions are split into two sections under the headings scenario 1 and scenario 2. In each case you are first presented with questions which involve you having to choose between two types of proposal: A and B. Proposals A and B only differ with respect to the levels of the criteria (defined over the page), all other factors remain the same. For every question in these sections we would like you to choose between A and B. Following this, you are asked to evaluate the importance of the criteria (as used in the priority-setting process) to you.

The criteria are set as below:

- **potential health gain**
- **evidence of clinical effectiveness**
- **budgetary impact**
- **equity of access and health status inequalities**
- **quality of service**
- **community values and priorities.**

(please turn over)

These can be further described in detail with their corresponding levels.

- **Potential health gain**

  This takes into account the number of people that are affected, the effect of the proposal (i.e. whether it is potentially a life-saving intervention or potentially improves quality of life), and the time span over which the health gain may occur. Note that health gain includes ill health avoided (i.e. prevention):

life-saving now	–	life-saving intervention for the majority (~ 80%) of those affected now
sustained improvement now	–	significant sustained improvement in physical/mental health for majority now
sustained improvement later	–	significant sustained improvement in majority later
temporary improvement now	–	temporary improvement in majority now
temporary improvement later	–	temporary improvement in majority later.

- **Evidence of clinical effectiveness**

  This is a measure of the quality of evidence used to support the clinical benefits/effectiveness and potential health gain for each of the proposals. In this exercise, it is assumed that the nature of the proposals mean that randomised controlled trials (RCTs) are possible. The levels have been based on those put forward by SIGN (Scottish Inter-Collegiate Guidelines Network) and are widely recognised by consultants, GPs and managers in the Scottish health service:

MA	–	evidence obtained from meta-analysis (MA) or RCTs
RCT	–	evidence obtained from at least one RCT
descriptive	–	evidence obtained from at least one well-conducted clinical study but no RCT (e.g. controlled study without randomisation, quasi-experimental, descriptive)
expert opinion	–	evidence obtained from expert committee reports or opinions and/or clinical experience of respected authorities
none	–	no evidence available.

- **Budgetary impact**

  Under this criterion, the additional cost that would be needed in order to implement the proposal is considered relative to the size of the current allocation to the specific disease/patient group in the examples (in this instance we are assuming a budget allocation of £1,000,000). Under this configuration, possible levels are:

big save	–	the proposal involves a cost saving (> £50,000)
small save	–	the proposal involves a cost saving (between £30,000 and £50,000)
none	–	no additional (or small additional) expense is required/saved, current resources are just reallocated within the disease area/patient group (between –£30,000 and +£30,000)
small expense	–	proposal involves small overall additional cost relative to what the disease area/patient group receives at present (between £30,000 and £50,000)
big expense	–	proposal involves large overall additional cost relative to what the disease area/patient group receives at present (> £50,000).

- **Equity of access and health status inequalities**

  Within these levels, the definition of 'deprived' is used to mean: 'those who do not have the same opportunity to use health services because of their ethnicity, social class, age etc.', and 'remote' refers to: 'those who do not have the same opportunity to access health services because of where they live'. Possible levels therefore are:

big reduction in inequality	–	proposal targets remote and deprived areas exclusively
small reduction in inequality	–	proposal targets either remote or deprived areas exclusively
remains the same	–	no differentiation by deprivation or remoteness
small increase in inequality	–	proposal targets only non-deprived areas or those that benefit most from the service at present
big increase in inequality	–	proposal targets only non-deprived and non-remote areas.

- **Quality of service**

  This refers to how the service is provided, which would be affected by the implementation of the proposal. In this case, the process quality is measured using the following criteria:

  – meets waiting time targets
  – conforms to local/national objectives (e.g. a move towards primary level care)
  – incorporates continuity of care
  – enables patients to be informed about their care.

  The levels are:

two or more direct 'hits'	–	the proposal improves the acceptability of the way the service is delivered by fully addressing two or more of the aspects of quality as defined above
one direct 'hit'	–	the proposal improves the acceptability of the way the service is delivered by fully addressing one of the aspects of quality as defined above
two or more partial 'hits'	–	the proposal improves the acceptability of the way the service is delivered by partially addressing two or more of the aspects of quality as defined above
one partial 'hit'	–	the proposal improves the acceptability of the way the service is delivered by partially addressing one of the aspects of quality as defined above
no 'hits'	–	the nature of the proposal means that none of these aspects of quality is addressed.

- **Community values and priorities**

  In this criterion, the opinions of the public, patients and carers, about the proposal, are taken into consideration. The levels reflect the opinion (i.e. whether the proposal was supported or objected to) and the quality of this information (i.e. the method used to collate these opinions):

  robust evidence 'support' – a collection of well-conducted studies or a research project that fully explores the opinions of community groups (public, patients, carers) using robust elicitation methods (e.g. jury mechanisms, several focus groups, large quality satisfaction surveys, or quantitative methods) which shows that community preferences support the proposal

  weak evidence 'support' – research using anecdotal evidence, small satisfaction surveys and small-scale quantitative surveys identify the preferences/opinions of some community members which indicates support for the proposal

  robust evidence 'indifferent' – evidence collected using robust elicitation methods (as defined above) which shows that the community is divided over whether they support or object to the proposal

  weak evidence 'object' – research using weaker evidence which indicates community objection to the proposal

  robust evidence 'object' – research using robust evidence which shows that community preferences object to the proposal.

Scenario 1

You have to imagine that you are comparing two proposals which have been put forward for implementation. They both address NHSiS [NHS in Scotland] priorities in Scotland or the equivalent NHS Lead Priorities in England and Wales; in particular, they are concentrated with the disease area of cancer (e.g. gynaecological cancer). The only way that the two proposals differ from each other is through the levels that are associated with the criteria (as defined above). You are now asked to make a choice between whether you prefer proposal A or B. Consider each choice separately and indicate your preference by ticking the appropriate box. Please tick one box for every choice: there are no right or wrong answers.

Section 1

• **Choice 1**	**Proposal A**	**Proposal B**
Potential health gain	Temporary improvement later	Life-saving now
Evidence of clinical effectiveness	RCT	MA
Budgetary impact	Small save	Small save
Equity of access and health status inequalities	Small increase in inequality	Remains the same
Quality of service	Two or more direct 'hits'	One partial 'hit'
Community values and priorities	Robust evidence 'indifferent'	Robust evidence 'object'

Which proposal would you prefer? (*tick one box only*)

Prefer proposal A ☐ Prefer proposal B ☐

• **Choice 2**	**Proposal A**	**Proposal B**
Potential health gain	Temporaryimprovement now	Temporary improvement later
Evidence of clinical effectiveness	None	Descriptive
Budgetary impact	Small expense	None
Equity of access and health status inequalities	Small increase in inequality	Big reduction in inequality
Quality of service	One partial 'hit'	One partial 'hit'
Community values and priorities	Weak evidence 'object'	Weak evidence 'support'

Which proposal would you prefer? (*tick one box only*)

Prefer proposal A ☐ Prefer proposal B ☐

• **Choice 3**	**Proposal A**	**Proposal B**
Potential health gain	Temporary improvement now	Life-saving now
Evidence of clinical effectiveness	Descriptive	Descriptive
Budgetary impact	Small save	Small expense
Equity of access and health status inequalities	Big increase in inequality	Small reduction in inequality
Quality of service	Two or more partial 'hits'	No 'hits'
Community values and priorities	Robust evidence in support	Robust evidence 'indifferent'

Which proposal would you prefer? (*tick one box only*)

Prefer proposal A ☐ Prefer proposal B ☐

• **Choice 4**	**Proposal A**	**Proposal B**
Potential health gain	Sustained improvement now	Life-saving now
Evidence of clinical effectiveness	None	None
Budgetary impact	Small save	Big save
Equity of access and health status inequalities	Small reduction in inequality	Big reduction in inequality
Quality of service	One direct 'hit'	Two or more direct 'hits'
Community values and priorities	Weak evidence 'support'	Robust evidence 'support'

Which proposal would you prefer? (*tick one box only*)

Prefer proposal A ☐ Prefer proposal B ☐

• **Choice 5**	**Proposal A**	**Proposal B**
Potential health gain	Life-saving now	Sustained improvement now
Evidence of clinical effectiveness	Expert opinion	RCT
Budgetary impact	Big expense	Small expense
Equity of access and health status inequalities	Small increase in inequality	Big reduction in inequality
Quality of service	Two or more partial 'hits'	Two or more partial 'hits'
Community values and priorities	Weak evidence 'support'	Robust evidence 'object'

Which proposal would you prefer? (*tick one box only*)

Prefer proposal A ☐ Prefer proposal B ☐

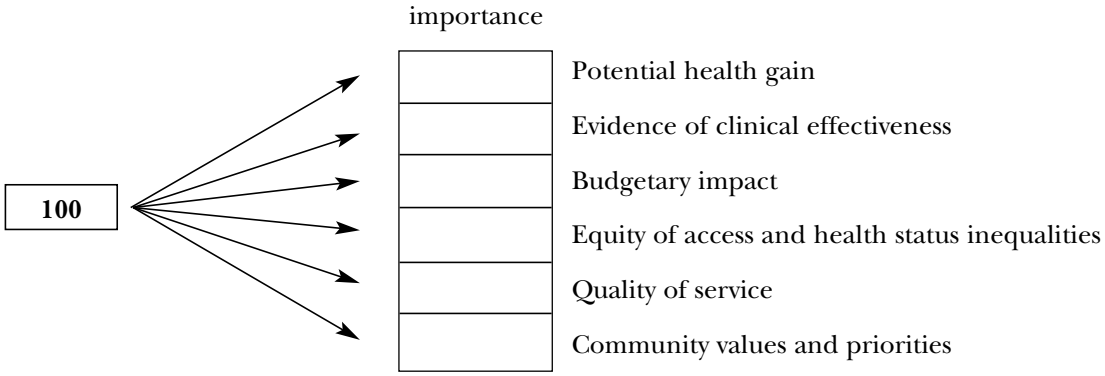
• **Choice 6**	**Proposal A**	**Proposal B**
Potential health gain	Temporary improvement now	Sustained improvement later
Evidence of clinical effectiveness	MA	Expert opinion
Budgetary impact	Big expense	Small save
Equity of access and health status inequalities	Big reduction in inequality	Big reduction in inequality
Quality of service	One direct 'hit'	No 'hits'
Community values and priorities	Robust evidence 'indifferent'	Weak evidence 'object'

Which proposal would you prefer? (*tick one box only*)

Prefer proposal A ☐ Prefer proposal B ☐

Section 2

In this case, you are asked to imagine (for scenario 1) that you have been given a fixed budget of 100 points, which you have to allocate to each of the criteria that have been described throughout this exercise:



You should allocate points to the criteria from your given budget. All of the budget can be allocated to one criterion if you so wish and not all of the criteria have to receive points. The number of points you give reflects the importance, in the priority-setting process, of that criterion to you. For example, if you allocate twice as many points to 'potential health gain' compared to 'quality of service' this means that you think that the potential health gain criterion is twice as important as the quality of service criterion when setting priorities.

Please ensure that before you leave this question you have allocated <u>all</u> 100 points and that the 'importance' column totals 100.

Scenario 2

You have to imagine that you are comparing two proposals which have been put forward for implementation. Neither address NHSiS [NHS in Scotland] priorities in Scotland or the equivalent NHS Lead Priorities in England and Wales; in this case, they are concerned with a potential but not 'big' killer (e.g. asthma). The only way that the two proposals differ from each other is through the levels that are associated with the criteria (as defined above). You are now asked to make a choice between whether you prefer proposal A or B. Consider each choice separately and indicate your preference by ticking the appropriate box. Please tick one box for every choice: there are no right or wrong answers.

Section 1

• **Choice 7**	**Proposal A**	**Proposal B**
Potential health gain	Sustained improvement later	Sustained improvement later
Evidence of clinical effectiveness	Descriptive	RCT
Budgetary impact	Big save	Big expense
Equity of access and health status inequalities	Small increase in inequality	Small reduction in inequality
Quality of service	One direct 'hit'	One partial 'hit'
Community values and priorities	Robust evidence 'object'	Robust evidence 'support'

Which proposal would you prefer? (*tick one box only*)

Prefer proposal A ☐ Prefer proposal B ☐

• **Choice 8**	**Proposal A**	**Proposal B**
Potential health gain	Sustained improvement later	Life-saving now
Evidence of clinical effectiveness	None	RCT
Budgetary impact	None	None
Equity of access and health status inequalities	Remains the same	Big increase in inequality
Quality of service	Two or more partial 'hits'	One direct 'hit'
Community values and priorities	Robust evidence 'indifferent'	Weak evidence 'object'

Which proposal would you prefer? (*tick one box only*)

Prefer proposal A ☐ Prefer proposal B ☐

• **Choice 9**	**Proposal A**	**Proposal B**
Potential health gain	Temporary improvement now	Temporary improvement later
Evidence of clinical effectiveness	Expert opinion	MA
Budgetary impact	None	Big save
Equity of access and health status inequalities	Small reduction in inequality	Small reduction in inequality
Quality of service	Two or more direct 'hits'	Two or more partial 'hits'
Community values and priorities	Robust evidence 'object'	Weak evidence object

Which proposal would you prefer? (*tick one box only*)

Prefer proposal A ☐ Prefer proposal B ☐

• **Choice 10**	**Proposal A**	**Proposal B**
Potential health gain	Sustained improvement now	Temporary improvement later
Evidence of clinical effectiveness	MA	Expert opinion
Budgetary impact	None	Small expense
Equity of access and health status inequalities	Small increase in inequality	Remains the same
Quality of service	Not relevant	One direct 'hit'
Community values and priorities	Robust evidence 'support'	Robust evidence 'support'

Which proposal would you prefer? (*tick one box only*)

Prefer proposal A ☐ Prefer proposal B ☐

• **Choice 11**	**Proposal A**	**Proposal B**
Potential health gain	Temporary improvement later	Sustained improvement now
Evidence of clinical effectiveness	None	Expert opinion
Budgetary impact	Big expense	Big save
Equity of access and health status inequalities	Big increase in inequality	Big increase in inequality
Quality of service	Not relevant	One partial 'hit'
Community values and priorities	Robust evidence 'object'	Robust evidence 'object'

Which proposal would you prefer? (*tick one box only*)

Prefer proposal A ☐ Prefer proposal B ☐

• **Choice 12**	**Proposal A**	**Proposal B**
Potential health gain	Sustained improvement now	Sustained improvement later
Evidence of clinical effectiveness	Descriptive	MA
Budgetary impact	Big expense	Small expense
Equity of access and health status inequalities	Remains the same	Big increase in inequality
Quality of service	Two or more direct 'hits'	Two or more direct 'hits'
Community values and priorities	Weak evidence 'object'	Weak evidence 'support'

Which proposal would you prefer? (*tick one box only*)

Prefer proposal A ☐ Prefer proposal B ☐

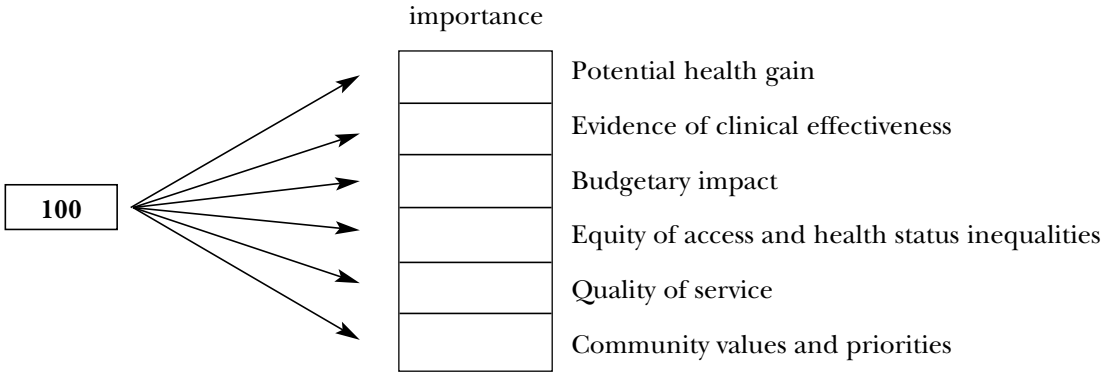
• **Choice 13**	**Proposal A**	**Proposal B**
Potential health gain	Temporary improvement now	Life-saving now
Evidence of clinical effectiveness	RCT	Expert opinion
Budgetary impact	Big save	Big expense
Equity of access and health status inequalities	Remains the same	Small increase in inequality
Quality of service	No 'hits'	Two or more partial 'hits'
Community values and priorities	Weak evidence 'support'	Weak evidence 'support'

Which proposal would you prefer? (*tick one box only*)

Prefer proposal A ☐ Prefer proposal B ☐

Section 2

In this case, you are asked to imagine (for scenario 2) that you have been given a fixed budget of 100 points, which you have to allocate to each of the criteria that have been described throughout this exercise:



You should allocate points to the criteria from your given budget. All of the budget can be allocated to one criterion if you so wish and not all of the criteria have to receive points. The number of points you give reflects the importance, in the priority-setting process, of that criterion to you. For example, if you allocate twice as many points to 'potential health gain' compared to 'quality of service' this means that you think that the potential health gain criterion is twice as important as the quality of service criterion when setting priorities.

Please ensure that before you leave this question you have allocated <u>all</u> 100 points and that the 'importance' column totals 100.

Finally, could you please provide a few details about yourself? The information from the following questions will not be used to identify individuals but to assist the analysis. **All answers will be treated as confidential.**

1. What is your age? ☐

2. What is your gender? Female ☐ Male ☐

3. What is your job title? _____

4. a. Do you have 'hands on experience' of priority setting? Yes ☐ No ☐

   b. If 'Yes', at what level (trust, health board/authority)?

   _____

   c. And for how long have you being doing this type of work?

   _____

5. Do you have experience of setting priorities using weighted criteria? Yes ☐ No ☐

6. a. Do you think that weighted criteria are a useful means by which to set priorities? Yes ☐ No ☐

   b. If 'No', give details:

7. Are there any other criteria that you would consider to be important for priority setting? (please list)

Thank you for taking the time to complete this questionnaire

If you would be willing to be contacted at a later date to discuss your views further please tick the relevant boxes:

I am willing to be contacted by telephone Yes ☐ No ☐

Telephone number: _____

The best time to contact me would be:

morning	
afternoon	
evening	
anytime	

Are there any comments you would like to make regarding the questionnaire?

Please return your questionnaire in the envelope provided

Appendix 7

Telephone interview schedule

Again, thank you for completing the questionnaire. I would now like to do some follow-up from the questionnaire and ask you some brief questions related to it.

The questionnaire in general

1. Do you think that you were a relevant recipient for this questionnaire?
   - Yes
   - No

The criteria

2. Looking at each of the criteria in turn: did you understand all of the descriptions and terminology (1 = didn't understand; 2 = could only understand a little; 3 = understood most; 4 = understood all)?
   a. Potential health gain
   b. Evidence of clinical effectiveness
   c. Budgetary impact
   d. Equity of access and health status inequalities
   e. Quality of service
   f. Community values and priorities

3. Were any too brief, i.e. not explained enough?
   - Yes: which?
   - No

4. Were the descriptions accurate or precise enough (1 = none was very precise or accurate; 2 = some were; 3 = most of them were; 4 = all very precise and accurate)?

5. Are there any additional criteria that you think were omitted from this section?

The scenarios

The scenarios were designed to incorporate other criteria that are of importance in decision-making, such as whether the proposal fulfilled a national priority/strategy or was concerned with a big disease area. Do you think that this was achieved?

- Yes
- No

6. Did you understand all of the terms used in the scenarios (1 = didn't understand any; 2 = could only understand some; 3 = understood most; 4 = understood all)?

7. Were they relevant?
   - Yes
   - No

8. Did you take them into account when answering the questions?
   - Yes
   - No

9. Were different criteria important under the two different scenarios?
   - Yes: which?
   - No

The paired choices

How easy was it to decide between the choices (on a scale of 1–4: 1 = very difficult; 4 = very easy)?

10. Did you find the hypothetical proposals too hypothetical or were they realistic (on a scale of 1–4: 1 = very hypothetical/unrealistic; 4 = not at all hypothetical/very realistic)?

11. Did anything appear as, in reality, contradictory?
    - Yes: what?
    - No

12. How did you decide between the choices? Did you either:
    - consider all the criteria and 'weigh-up' the alternatives? Or
    - only concentrate on some criteria that you thought were important? What were these?

13. Was anything irrelevant or did some criteria become redundant when making your choices?
    - Yes: what?
    - No

14. Were there any choices that were particularly hard to make?
    - Yes: what?
    - No

The budget allocation question

How easy was it to divide points between the various criteria (on a scale of 1–4: 1 = very difficult; 4 = very easy)?

15. How did you divide the points between the criteria (top, middle, bottom, all allocated the same points for example)?

16. Do you think that the number of points you gave something accurately reflected your strength of preference for that criteria, i.e. if you gave one 20 points and another 40 points, did this mean the latter was twice as important to you?
    - Yes
    - No: why?

Comparing the two methods

17. Was either method – paired choice or budget allocation:
    - easier – which?
    - quicker – which?

The results

The results have shown that differences exist between the results collected using the two different methods. To test these methods could you:

18. Rank in order of importance in terms of priority decision-making (in your view) the criteria described in the questionnaire?

19. Of the least important criteria, are any of them in fact irrelevant or not important at all – which?

20. Would you consider any of the criteria to be of equal importance (i.e. ranked the same) – which?

Priority setting in general

21. Do you think that patients or the community have a role to play in priority setting?
    - Yes: what?
    - No: why?

22. Do you think that the views of the public or patients are important when deciding whether or not to implement a proposal (on a scale of 1–4: 1 = not at all important; 4 = very important)?

Future work

Future work is planned to investigate further the importance of these criteria in priority-setting decision-making contexts.

23. Do you have any comments for the future study?

Methodology Group

Members

Methodology Programme Director
Professor Richard Lilford
Director of Research and Development
NHS Executive – West Midlands, Birmingham

Chair
Professor Martin Buxton
Director, Health Economics Research Group
Brunel University, Uxbridge

Professor Douglas Altman
Professor of Statistics in Medicine
University of Oxford

Dr David Armstrong
Reader in Sociology as Applied to Medicine
King's College, London

Professor Nicholas Black
Professor of Health Services Research
London School of Hygiene & Tropical Medicine

Professor Ann Bowling
Professor of Health Services Research
University College London Medical School

Professor David Chadwick
Professor of Neurology
The Walton Centre for Neurology & Neurosurgery
Liverpool

Dr Mike Clarke
Associate Director (Research)
UK Cochrane Centre, Oxford

Professor Paul Dieppe
Director, MRC Health Services Research Centre
University of Bristol

Professor Michael Drummond
Director, Centre for Health Economics
University of York

Dr Vikki Entwistle
Senior Research Fellow, Health Services Research Unit
University of Aberdeen

Professor Ewan B Ferlie
Professor of Public Services Management
Imperial College, London

Professor Ray Fitzpatrick
Professor of Public Health & Primary Care
University of Oxford

Dr Naomi Fulop
Deputy Director, Service Delivery & Organisation Programme
London School of Hygiene & Tropical Medicine

Mrs Jenny Griffin
Head, Policy Research Programme
Department of Health
London

Professor Jeremy Grimshaw
Programme Director
Health Services Research Unit
University of Aberdeen

Professor Stephen Harrison
Professor of Social Policy
University of Manchester

Mr John Henderson
Economic Advisor
Department of Health, London

Professor Theresa Marteau
Director, Psychology & Genetics Research Group
Guy's, King's & St Thomas's School of Medicine, London

Dr Henry McQuay
Clinical Reader in Pain Relief
University of Oxford

Dr Nick Payne
Consultant Senior Lecturer in Public Health Medicine
ScHARR
University of Sheffield

Professor Joy Townsend
Director, Centre for Research in Primary & Community Care
University of Hertfordshire

Professor Kent Woods
Director, NHS HTA Programme, & Professor of Therapeutics
University of Leicester

continued

HTA Commissioning Board

Members

Programme Director
Professor Kent Woods
Director, NHS HTA Programme, &
Professor of Therapeutics
University of Leicester

Chair
Professor Shah Ebrahim
Professor of Epidemiology of Ageing
University of Bristol

Deputy Chair
Professor Jon Nicholl
Director, Medical Care Research Unit
University of Sheffield

Professor Douglas Altman
Director, ICRF Medical Statistics Group
University of Oxford

Professor John Bond
Director, Centre for Health Services Research
University of Newcastle-upon-Tyne

Ms Christine Clark
Freelance Medical Writer
Bury, Lancs

Professor Martin Eccles
Professor of Clinical Effectiveness
University of Newcastle-upon-Tyne

Dr Andrew Farmer
General Practitioner &
NHS R&D Clinical Scientist
Institute of Health Sciences
University of Oxford

Professor Adrian Grant
Director, Health Services Research Unit
University of Aberdeen

Dr Alastair Gray
Director, Health Economics Research Centre
Institute of Health Sciences
University of Oxford

Professor Mark Haggard
Director, MRC Institute of Hearing Research
University of Nottingham

Professor Jenny Hewison
Senior Lecturer
School of Psychology
University of Leeds

Professor Alison Kitson
Director, Royal College of Nursing Institute, London

Dr Donna Lamping
Head, Health Services Research Unit
London School of Hygiene & Tropical Medicine

Professor David Neal
Professor of Surgery
University of Newcastle-upon-Tyne

Professor Gillian Parker
Nuffield Professor of Community Care
University of Leicester

Dr Tim Peters
Reader in Medical Statistics
University of Bristol

Professor Martin Severs
Professor in Elderly Health Care
University of Portsmouth

Dr Sarah Stewart-Brown
Director, Health Services Research Unit
University of Oxford

Professor Ala Szczepura
Director, Centre for Health Services Studies
University of Warwick

Dr Gillian Vivian
Consultant in Nuclear Medicine & Radiology
Royal Cornwall Hospitals Trust
Truro

Professor Graham Watt
Department of General Practice
University of Glasgow

Dr Jeremy Wyatt
Senior Fellow
Health Knowledge Management Centre
University College London

Feedback

The HTA programme and the authors would like to know your views about this report.

The Correspondence Page on the HTA website (http://www.ncchta.org) is a convenient way to publish your comments. If you prefer, you can send your comments to the address below, telling us whether you would like us to transfer them to the website.

We look forward to hearing from you.

Copies of this report can be obtained from:

The National Coordinating Centre for Health Technology Assessment,
Mailpoint 728, Boldrewood,
University of Southampton,
Southampton, SO16 7PX, UK.
Fax: +44 (0) 23 8059 5639 Email: hta@soton.ac.uk
http://www.ncchta.org

ISSN 1366-5278