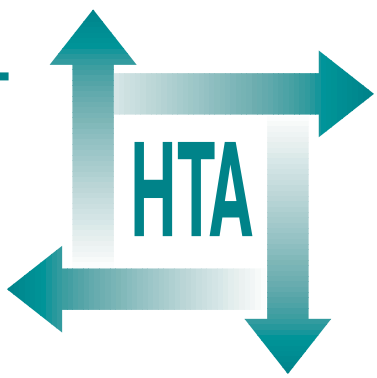
Health Technology Assessment 1998; Vol. 2: No. 3

Review

Consensus development methods, and their use in clinical guideline development

MK Murphy
NA Black
DL Lamping
CM McKee
CFB Sanderson
J Askham
T Marteau

Health Technology Assessment
NHS R&D HTA Programme



Standing Group on Health Technology

Chair: Professor Sir Miles Irving,
Professor of Surgery, University of Manchester, Hope Hospital, Salford †

Dr Sheila Adam,
Department of Health

Professor Martin Buxton,
Professor of Economics, Brunel University †

Professor Angela Coulter,
Director, King's Fund, London

Professor Anthony Culyer,
Deputy Vice-Chancellor, University of York

Dr Peter Doyle,
Executive Director, Zeneca Ltd,
ACOST Committee on Medical Research & Health

Professor John Farndon,
Professor of Surgery, University of Bristol †

Professor Charles Florey,
Department of Epidemiology & Public Health, Ninewells Hospital & Medical School, University of Dundee †

Professor John Gabbay,
Director, Wessex Institute for Health Research & Development †

Dr Tony Hope,
The Medical School, University of Oxford †

Professor Howard Glennester,
Professor of Social Science & Administration, London School of Economics & Political Science

Professor Sir John Grimley Evans,
Department of Geriatric Medicine, Radcliffe Infirmary, Oxford †

Mr John H James,
Chief Executive, Kensington, Chelsea & Westminster Health Authority

Professor Richard Lilford,
Regional Director, R&D, West Midlands †

Professor Michael Maisey,
Professor of Radiological Sciences, UMDS, London

Dr Jeremy Metters,
Deputy Chief Medical Officer, Department of Health †

Mrs Gloria Oates,
Chief Executive, Oldham NHS Trust

Dr George Poste,
Chief Science & Technology Officer, SmithKline Beecham †

Professor Michael Rawlins,
Wolfson Unit of Clinical Pharmacology, University of Newcastle-upon-Tyne

Professor Martin Roland,
Professor of General Practice, University of Manchester

Mr Hugh Ross,
Chief Executive, The United Bristol Healthcare NHS Trust †

Professor Ian Russell,
Department of Health, Sciences & Clinical Evaluation, University of York

Professor Trevor Sheldon,
Director, NHS Centre for Reviews & Dissemination, University of York †

Professor Mike Smith,
Director, The Research School of Medicine, University of Leeds †

Dr Charles Swan,
Consultant Gastroenterologist, North Staffordshire Royal Infirmary

Dr John Tripp,
Department of Child Health, Royal Devon & Exeter Healthcare NHS Trust †

Professor Tom Walley,
Department of Pharmacological Therapeutics, University of Liverpool †

Dr Julie Woodin,
Chief Executive, Nottingham Health Authority †

† Current members

HTA Commissioning Board

Chair: Professor Charles Florey, Department of Epidemiology & Public Health,
Ninewells Hospital & Medical School, University of Dundee †

Professor Ian Russell,
Department of Health, Sciences & Clinical Evaluation, University of York *

Dr Doug Altman,
Director, Institute of Health Sciences, Oxford †

Mr Peter Bower,
Independent Management Consultant, Newcastle-upon-Tyne †

Ms Christine Clark,
Hon. Research Pharmacist, Hope Hospital, Salford †

Professor David Cohen,
Professor of Health Economics, University of Glamorgan

Mr Barrie Dowdeswell,
Chief Executive, Royal Victoria Infirmary, Newcastle-upon-Tyne

Professor Martin Eccles,
Professor of Clinical Effectiveness, University of Newcastle-upon-Tyne †

Dr Mike Gill,
Brent & Harrow Health Authority †

Dr Jenny Hewison,
Senior Lecturer, Department of Psychology, University of Leeds †

Dr Michael Horlington,
Head of Corporate Licensing, Smith & Nephew Group Research Centre

Professor Sir Miles Irving
(Programme Director), Professor of Surgery, University of Manchester, Hope Hospital, Salford †

Professor Alison Kitson,
Director, Royal College of Nursing Institute †

Professor Martin Knapp,
Director, Personal Social Services Research Unit, London School of Economics & Political Science

Dr Donna Lamping,
London School of Hygiene & Tropical Medicine †

Professor Theresa Marteau,
Director, Psychology & Genetics Research Group, UMDS, London

Professor Alan Maynard,
Professor of Economics, University of York †

Professor Sally McIntyre,
MRC Medical Sociology Unit, Glasgow

Professor Jon Nicholl,
Director, Medical Care Research Unit, University of Sheffield †

Professor Gillian Parker,
Nuffield Professor of Community Care, University of Leicester †

Dr Tim Peters,
Department of Social Medicine, University of Bristol †

Professor David Sackett,
Centre for Evidence Based Medicine, Oxford

Professor Martin Severs,
Professor in Elderly Health Care, Portsmouth University †

Dr David Spiegelhalter,
MRC Biostatistics Unit, Institute of Public Health, Cambridge

Dr Ala Szczepura,
Director, Centre for Health Services Studies, University of Warwick †

Professor Graham Watt,
Department of General Practice, Woodside Health Centre, Glasgow †

Professor David Williams,
Department of Clinical Engineering, University of Liverpool

Dr Mark Williams,
Public Health Physician, Bristol

Dr Jeremy Wyatt,
Institute for Health Sciences, University College London †

* Previous Chair
† Current members

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.

Consensus development methods, and their use in clinical guideline development

MK Murphy[1]
NA Black[1]
DL Lamping[1]
CM McKee[1]
CFB Sanderson[1]
J Askham[2]
T Marteau[3]

[1] Health Services Research Unit, London School of Hygiene & Tropical Medicine
[2] King's College, London
[3] United Medical & Dental Schools, London

Published March 1998

This report should be referenced as follows:

Murphy MK, Black NA, Lamping DL, McKee CM, Sanderson CFB, Askham J, *et al.* Consensus development methods, and their use in clinical guideline development. *Health Technol Assessment* 1998; **2**(3).

Health Technology Assessment is indexed in Index Medicus/Medline and Excerpta Medica/Embase. Copies of the Executive Summaries are available from the NCCHTA web site (see overleaf).

NHS R&D HTA Programme

The overall aim of the NHS R&D Health Technology Assessment (HTA) programme is 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 work in the NHS. Research is undertaken in those areas where the evidence will lead to the greatest benefits to patients, either through improved patient outcomes or the most efficient use of NHS resources.

The Standing Group on Health Technology advises on national priorities for health technology assessment. Six advisory panels assist the Standing Group in identifying and prioritising projects. These priorities are then considered by the HTA Commissioning Board supported by the National Coordinating Centre for HTA (NCCHTA).

This report is one of a series covering acute care, diagnostics and imaging, methodology, pharmaceuticals, population screening, and primary and community care. It was identified as a priority by the Methodology Panel.

The views expressed in this publication are those of the authors and not necessarily those of the Standing Group, the Commissioning Board, the Panel members 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 the recommendations for policy contained herein. In particular, policy options in the area of screening will, in England, be considered by the National Screening Committee. This Committee, chaired by the Chief Medical Officer, will take into account the views expressed here, further available evidence and other relevant considerations.

Series Editors: Andrew Stevens, Ruairidh Milne and Ken Stein
Assistant Editor: Jane Robertson

The editors have tried to ensure the accuracy of this report but cannot accept responsibility for any errors or omissions. They would like to thank the referees for their constructive comments on the draft document.

ISSN 1366-5278

© Crown copyright 1998

Enquiries relating to copyright should be addressed to the NCCHTA (see address given below).

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

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) 1703 595 639 Email: hta@soton.ac.uk
http://www.soton.ac.uk/~hta

Contents

List of abbreviations

AAA	abdominal aortic aneurysm
ARR	absolute risk reduction*
CABG	coronary artery bypass graft
CI	confidence interval*
CDP	Consensus Development Panel*
CT scan	computed tomography scan
DA	decision analysis*
EFP	event-free patients*
GP	general practitioner
ICC	intra-class correlation coefficient
NGT	nominal group technique
NNT	number needed to treat*
PCL	problem centred leadership
PTCA	percutaneous transluminal coronary angioplasty
QALY	quality adjusted life-year
RRR	relative risk reduction*
SD	standard deviation*
SJA	social judgement analysis
TI	theoretical indication*

* Used only in tables

Executive summary

Background

Consensus methods are increasingly being used to develop clinical guidelines which define key aspects of the quality of health care, particularly appropriate indications for interventions. This review is restricted to formal consensus methods in which the structure, process and output are explicit from the outset. Three main approaches have been used in the health field: the Delphi method, the nominal group technique (NGT) and the consensus development conference.

Objectives

- To identify the factors that affect the decisions that emerge from consensus development methods.
- To assess the implications of the findings for the development of clinical guidelines.
- To recommend further methodological research for improving the use of consensus development methods as a basis for guideline production.

Methods

Data sources

The majority of the literature reviewed was identified through searches of Medline, PsychLIT and the Social Science Citation Index and from reference lists in retrieved articles.

Study selection

A matrix of 15 cells was developed from three types of activity (planning, individual judgement, group interaction) and five components (questions, participants, information, method of structuring the interaction, method of synthesising individual judgements) involved in consensus development methods.

Six cells were selected for detailed review on the basis of three criteria: (1) importance to consensus decision-making in the health sector; (2) the amount and quality of the literature available; (3) the potential for offering practical guidance. For each of the six cells the review drew on the results of the principal general search. For some cells, further focused searches were undertaken. In all, 177 primary research and review articles were selected.

Data extraction and synthesis

If substantial literature was available from the health sector, we paid little or no attention to evidence from other sectors. If few or no studies had been conducted in the health sector, we sought relevant evidence from other fields. We used a narrative approach, sometimes based around tables of results. The extent to which research support exists for any conclusion is indicated, although these should not necessarily be considered as a hierarchy: A = clear research evidence; B = limited supporting research evidence; C = experienced common-sense judgement.

Results and conclusions

Setting the task or question to be addressed

- Cues included in scenarios must be selected with care. As well as reviewing the relevant literature, clinicians in the consensus group should give their opinions (most usefully in the first round) about which cues are important. Doing so may help maintain their participation and help them justify their judgements. [C]
- Contextual cues included in scenarios are as important as ones specific to the topic at issue, and they should be made explicit. [B]
- It must be decided whether to focus on ways of managing a specific condition or on indications for using an intervention. If the focus is on an intervention, care should be taken about how to deal with other relevant interventions. [C]
- Is a global judgement elicited, or is an attempt made to break the judgement down into probability and utility estimates? Although there are theoretical advantages to the latter, it is likely to be a more difficult task for participants and it may not enhance judgements. [C]
- Inclusion of all possible scenarios may increase comprehensiveness, but if many of the scenarios never occur in practice, the increased burden on the respondents may not be justified by the

limited value of the information provided. Judgements of scenarios which never or rarely occur in practice may be less reliable. [B]
- Requiring participants to judge what may be seen as numerous irrelevant scenarios may alienate them from the task. [C]

Selecting the participants

- Within defined specialist or professional categories, the selection of the particular individuals is likely to have little impact on the decision of a group of sufficient size. To enhance the credibility and widespread acceptance of the guidelines, the participants should reflect the full range of key characteristics of the population that it is intended to influence. Selection should be seen to be unbiased. [C]
- To define common ground and maximise areas of agreement, groups should be homogeneous; to identify and explore areas of uncertainty, a heterogeneous group is appropriate. [B]
- In judgements of clinical appropriateness, the most influential background factor is the particular medical specialty. Specialists tend to favour the interventions with which they are most familiar. Consensus-based guidelines should therefore be interpreted in the context of the specialty composition of the group. [A]

Choosing and preparing the scientific evidence

- A review of research-based information should be provided to all participants at an early stage. Participants should be encouraged to bring the review and any personal notes to the group sessions as memory aids. [B]
- Information presented in a synthesised form (e.g. tables) is more likely to be assimilated. Participants may be more likely to use information that is presented in an accessible format. Information tabulated so as to increase the salience of the dimensions to be used for making judgements is more likely to be processed in this manner. [C]
- Methodologists should be involved in conducting any literature review. [C]
- Grading the quality of studies using a reliable method may mitigate the biases of the reviewers somewhat, but may not eliminate them. [B]

Structuring the interaction

- With NGTs and the Delphi method, two or more rating rounds are likely to result in some convergence of individual judgements, though it is unclear whether this increases the accuracy of the group decision. [A]
- With the Delphi method, it is advisable to feed back reasons or arguments as well as measures of central tendency or dispersion. [B]
- Efforts should be made to mitigate the effects of status of participants (which can affect their contribution to and influence within a group). [B]
- A comfortable environment for meetings is likely to be preferred by participants and to be conducive to discussion. [C]
- A good facilitator will enhance consensus development and can ensure that the procedure is conducted properly. [C]

Methods of synthesising individual judgements

- An implicit approach to aggregating individual judgements may be adequate for establishing broad policy guidelines. More explicit methods based on quantitative analysis are needed to develop detailed, specific guidelines. [C]
- The more demanding the definition of agreement, the more anodyne the results will be. If the requirement is too demanding, either no statements will qualify or those that do will be of little interest. [C]
- Differential weighting of individual participants' views produces unreliable results unless there is a clear empirical basis for calculating the weights. [B]
- The exclusion of individuals with extreme views (outliers) can have a marked effect on the content of guidelines. [A]
- There is no agreement as to the best method of mathematical aggregation. [B]
- Reports of consensus development exercises should include an indication of the distribution or dispersal of participants' judgements, not just the measure of central tendency. In general, the median and the inter-quartile range are more robust than the mean and standard deviation. [A]

Priorities for future research

- What impact does the framing or presentation of the question have on individual judgements?
- In what form and how inclusive should scenarios be?
- How does the extent of heterogeneity of a group affect the process and outcome?
- What effect does research-based information have on individual and on group judgements? Does the effect depend on the amount of information or how it is presented?
- What effect does the method of feedback of participants' views have on group judgement?

Chapter 1
Introduction

Why we need consensus development methods

Clinicians regularly make difficult choices about treatment options. Often there is uncertainty about the value of different options, and practice can vary widely. Although there is debate about the appropriate place of guidelines in clinical practice, guidelines can be seen as one way of assisting clinicians in decision-making.

In an ideal world, clinical guidelines would be based on evidence derived from rigorously conducted empirical studies. In practice, there are few areas of health care where sufficient research-based evidence exists or may ever exist (Chassin, 1989). In such situations, the development of guidelines will inevitably have to be based partly or largely on the opinions and experience of clinicians and others with knowledge of the subject at issue (Agency for Health Care Policy and Research, 1995; Mann 1996).

There are two main ways in which judgement-based guidelines could be devised: have the 'best person' make the judgement, or have a group do it. Problems with the 'best person' model are that (1) it simply pushes the problem upstream to 'what is the best way of identifying the best person?', (2) no one person may have access to all the relevant information, (3) it may be a rather erratic and unsafe approach, and (4) the best person may have limited credibility, which may be a particular problem when the results need to be persuasive to have the desired effect (as in clinical guidelines).

In theory, there are a number of advantages of a group decision: a wider range of direct knowledge and experience is brought to bear; the interaction between members stimulates consideration of a wide range of options and debate that challenges received ideas and stimulates new ones; idiosyncrasies are filtered out (sometimes wrongly!); and, in terms of influencing the behaviour of others, the group as a whole may carry more weight than any one individual. However, there are several issues to be addressed: the choice of participants; how to avoid one or more individuals dominating the proceedings; the cost of bringing people together; and the tendency to treat group decisions as unanimous when the degree of dissent within the group is an important piece of information.

Given the likely diversity of opinion that any group of people may display when considering a topic, methods are needed for organising subjective judgements. Although various methods exist (which are described below), they share the common objective of synthesising judgements when a state of uncertainty (differences of opinion) exists.

Our concern is with the use of these methods for developing clinical guidelines, but they have been used for several other purposes in the health sector including forecasting, conflict resolution and prioritisation. Recently their use has been extended to exploring moral and ethical issues in health care (Stewart *et al,* 1994). There has also been a tendency, fuelled largely by agencies seeking to contain healthcare costs, to use the methods for rationing care. In addition, they have been used not to identify areas of agreement but rather to establish those areas where there is a lack of agreement. This may, for example, be the objective of research bodies wanting to identify potential areas for new research (Bond & Bond, 1982).

Despite their widespread use, consensus development methods have been the subject of relatively little methodological research within the health field (Black, 1994). They have, however, been the subject of a considerable amount of investigation elsewhere, in particular in the behavioural science, technological and social forecasting literature (Parente & Anderson-Parente, 1987; Rowe *et al,* 1991), but this research has had little impact on their application in health care.

It is essential to be clear about what consensus development is and what it is not. It is a process for making policy decisions, not a scientific method for creating new knowledge. At its best, consensus development merely makes the best use of available information, be that scientific data or the collective wisdom of the participants. Thus, although it may capture collective knowledge, it is inevitably vulnerable to the possibility of capturing collective ignorance. Enthusiasts

should also recognise the limited role of such methods because, unfortunately, they rarely resolve disputes where strong disagreement exists (Fletcher, 1997).

The present situation is that despite many unresolved methodological questions, consensus methods are increasingly being used to determine clinical guidelines which define key aspects of the quality of health care, in particular the appropriate indications for using interventions (investigations or treatments). Indeed, consensus development methods have recently been viewed not with reluctance but with enthusiasm in an editorial in *The Lancet* (Lancet, 1997).

The objectives of this review are:

- to identify the factors that shape and influence the decisions that emerge from consensus development methods, particularly as regards the development of clinical guidelines
- to make recommendations about best practice in the use of consensus development methods for producing clinical guidelines
- to recommend further methodological research for improving the use of consensus development methods as a basis for guideline production.

Why we need 'formal' methods of consensus development

It is only since the 1950s that formal consensus development methods have been used in the health sector. This does not mean that collective decisions were not made before then, simply that such decisions emerged through informal methods. Indeed, over the past 40 years the vast majority of collective decisions in health care have continued to be based on group meetings, such as committees, which have been largely unstructured with few formal rules or procedures.

Such group discussions, sometimes termed 'free discussion', 'freely interacting' or simply 'consensus' groups involve bringing together a group of people to discuss a problem with the aim of reaching agreement. They are usually not instructed on how to reach a consensus though they may be given simple instructions such as not to criticise other members' contributions. There may or may not be someone chairing the group. A jury is an example of this type of group.

Given the widespread use of informal methods for reaching group decisions, why bother with formal consensus methods? The main reason is that the presence of others has been shown to affect performance in a variety of ways, not always positive and beneficial. For example, early research on social facilitation showed that the performance of well-learned tasks can be improved when others are present whereas the performance of less well-mastered tasks can be inhibited (Zajonc, 1965). Research on conformity shows that people will sometimes conform to the judgements of others, even in their judgements of the lengths of lines (Asch, 1956). Conformity can occur for a variety of reasons. Group members may feel pressure, real or imagined, to say what others say, even if they know it is against their better judgement. People also use others to interpret the nature of reality, especially when that reality is ambiguous (Sherif, 1937). Thus, the presence and actions of others are far from being neutral in their effects.

The processes involved in groups may hinder decision-making. For example, a simple problem is that only one individual can speak at a time. This can limit the number of ideas discussed in a group (Diehl & Stroebe, 1987). The desire to reach agreement may override concerns about the accuracy of the result to the extent that there is premature closure on a particular solution without consideration of alternatives (Janis, 1982). Social pressures may also have a damaging effect on group performance. Domination of the discussion by particular individuals, or pressures to agree with a majority or powerful person's viewpoint, have all been suggested as reasons why groups can produce poorer results than individuals. Formal methods have been developed with the aim of overcoming some of these problems.

The case for using formal methods is therefore based on a number of assumptions about decision-making in groups:

- safety in numbers – several people are less likely to arrive at a wrong decision than a single individual
- authority – a selected group of individuals is more likely to lend some authority to the decision produced
- rationality – decisions are improved by reasoned argument in which assumptions are challenged and members forced to justify their views
- controlled process – by providing a structured process formal methods can eliminate negative aspects of group decision-making
- scientific credibility – formal consensus methods meet the requirements of scientific methods.

Types of formal consensus development methods

This review is restricted to a consideration of formal methods. The distinction between formal and informal consensus methods may not always be clear. While formal consensus methods follow agreed procedures, some of these techniques may incorporate an element of informal, free interaction.

Three main approaches have been used in the health field. In the 1950s the Delphi method was introduced (Dalkey & Helmer, 1963; Pill, 1971); this was followed by the use of the nominal group technique (NGT) in the 1960s (Delbecq & Van de Ven, 1971); and in 1977, the National Institute of Health in the USA introduced the consensus development conference (Fink *et al*, 1984). The major differences between these methods relate to:

- whether a mailed questionnaire is used
- whether individuals make separate decisions 'in private' or not, and if so, the degree of confidentiality
- whether information on the group's deliberations or interim decisions is fed back to the participants for reconsideration during the process
- whether there is face-to-face contact between group members, and if so, whether or not it is structured
- the method used to aggregate participants' views.

Table 1 summarises the characteristics of both informal and formal methods. Some of the differences arise from differing assumptions and aims. Unlike the NGT and the Delphi method, consensus development conferences were developed with the additional aim of providing a public forum for the discussion of issues. In contrast, the Delphi method and the NGT are concerned largely with deriving a group decision from a set of 'expert' individuals.

One feature common to all the methods, when used as a basis for creating clinical guidelines, is the use of cues. Cues are the dimensions or indications that group members are asked to take into account when making their decisions. For example, if participants were deciding on the appropriate use of a treatment, one of the cues they would need to consider would be the severity of the condition being treated. Others cues might include age, gender and co-morbidity. Some methods present cues to participants as part of a scenario or vignette – a description of a situation. Participants are presented with a set of scenarios, each describing a different clinical situation, and are asked to decide on the appropriateness of a particular intervention (investigation or treatment) in each.

There is a spectrum of methods for aggregating the judgements of individuals which can be characterised by the extent to which the method is implicit or explicit. Implicit methods tend to be qualitative or involve simple quantitative techniques (such as a majority vote). Explicit methods tend to be more complex, involving statistical methods in which judgements are combined

TABLE 1 *Characteristics of informal and formal consensus development methods*

Consensus development method	Mailed questionnaires	Private decisions elicited	Formal feedback of group choices	Face-to-face contact	Interaction structured	Aggregation method
Informal	No	No	No	Yes	No	Implicit
Delphi method	Yes	Yes	Yes	No	Yes	Explicit
NGT	No	Yes	Yes	Yes	Yes	Explicit
RAND version	Yes	Yes	Yes	Yes	Yes	Explicit
Consensus development conference	No	No	No	Yes	No	Implicit
Other methods						
Staticised group	No	Yes	No	No	–	Explicit
Social judgement analysis	No	Yes	Yes	Yes	No	Implicit
Structured discussion	No	No	No	Yes	Yes	Implicit

according to mathematical rules, for example by taking the mean of individual judgements. Methods such as consensus development conferences rely on implicit methods whereas the Delphi method and the NGT use explicit, mathematical integration.

Delphi method

Participants never meet or interact directly. Instead, they are sent questionnaires and asked to record their views. Commonly, participants are initially asked to suggest the factors or cues that should be considered by the group. Having contributed to drawing up the agenda, in the next stage the participants are sent a questionnaire which seeks their individual views about the items that they and their co-participants have suggested. The responses are collated by the organisers and sent back to participants in summary form, usually indicating the group judgement and the individual's initial judgement. Participants are given the opportunity to revise their judgement in the light of the group feedback. This process may be repeated a number of times. The judgements of participants are then statistically aggregated, sometimes after weighting for expertise.

The Delphi method was developed by the RAND Corporation in the 1950s. It was originally used in forecasting – as reflected by its naming after the Greek oracle at Delphi which was believed to have the power to predict the future (but which was notorious for the ambiguity of its utterances!). The aim of the RAND Corporation was to synthesise expert opinion, mainly on the emergence of new technologies. Since then, the Delphi method has been used for a variety of purposes in the health sector (*Table 2*), though rarely for developing clinical guidelines.

The logic behind the Delphi method is partly statistical – combined numerical estimates of participants' views would, in general, lead to more reliable estimates than estimates from a single person. In addition, the Delphi method allows information to be exchanged between individuals (who may be numerous and geographically dispersed) in an iterative process, in the belief that there will be benefits from the exchange of information at low cost. However, this exchange is strictly controlled to limit the potentially detrimental effects of interaction. Conversely, the method has been criticised for diminishing the potentially positive aspects of interaction to be found in the face-to-face exchange of information which helps identify the reasons for any disagreements.

TABLE 2 *Examples of applications of the Delphi method in the health field*

Study	Reference
Research priorities for trauma nursing	Bayley *et al,* 1994
Clinical nursing research priorities	Bond & Bond, 1982
Research and service priorities in primary health care for persons with physical disabilities	Burns *et al,* 1990
Strategies for surviving cutbacks in community mental health programmes	Goplerud *et al,* 1985
Priorities and recommendations concerning patient education carried out by the GP	Grol *et al,* 1991
Agenda for clinical nursing research in long-term care	Haight & Bahr, 1992
Priorities for research in occupational medicine	Harrington, 1994
Effects on quality of care of reductions in number of junior doctors	Jones *et al,* 1992
Agreement on classifications of electrocardiograms	Kors *et al,* 1990
Development of a malignant hyperthermia clinical grading scale	Larach *et al,* 1994
Defining characteristics of ineffective breastfeeding	Lethbridge *et al,* 1993
Validation of definitions and activities important to techniques of pain management	Mobily *et al,* 1993
Appropriate preventive therapy for contacts of multidrug-resistant tuberculosis	Passannante *et al,* 1994
Forecasting the future of the hospital pharmacy	Plumridge, 1981
Agreeing terminology for substance abuse	Rinaldi *et al,* 1988

Nominal group technique

The NGT was developed by Delbecq and Van de Ven (1971) in the context of committee decision-making. They saw the non-interacting aspect of NGTs as suitable for situations involving individuals with differing views in which one objective was to encourage the generation of ideas. It has, however, also been used for other tasks including the development of clinical guidelines (*Table 3*).

The aim of NGTs is to structure interaction within a group. Firstly, each participant records his or her ideas independently and privately. The ideas are then listed in a round-robin format, that is one idea is collected from each individual in turn and listed in front of the group by the facilitator, and the process is continued until all ideas have been listed. Each idea is then discussed in turn by the group. Individuals then privately record their judgements or vote for options. Further discussion and voting may take place. The individual judgements are aggregated statistically to derive the group judgement.

In practice, formal consensus methods often involve variations on the techniques described. This is particularly true for NGTs. The most commonly used method for clinical guideline production is a 'modified NGT' developed by the RAND Corporation during the 1970s and 1980s (Bernstein *et al,* 1992), although the developers referred to it as a 'modified Delphi'. Initially individuals express their views privately via mailed questionnaires. The collated results of the questionnaire are fed back to each member of the group when they are brought together to discuss their views, after which they again privately record their views on a questionnaire (see the box on page 6).

Delbecq and Van de Ven (1971) suggested that the NGT would be superior to informal groups for two reasons. Firstly, informal groups can inhibit members from speaking freely and sharing what may be under-developed ideas. Secondly, informal groups may focus too much on one particular idea, often one generated early on, and fail to explore the problem thoroughly. NGTs avoid evaluation and elaboration of ideas during the phase of their generation, thereby allowing more ideas to be expressed and elaborated. The developers believed that each person would be more likely to work on the problem, rather than leaving the generation of ideas to someone else ('social loafing' or 'free riding'). In an NGT, therefore, each person is more likely to generate ideas and to be less inhibited

TABLE 3 *Examples of NGT applications in the health field*

Study	Reference
Strategies for implementing new nursing practices	Buchan *et al,* 1991
Priorities for health promotion	Brown & Redman, 1995
Development of a set of disease activity measures for use in rheumatoid arthritis	Felson *et al,* 1993
Identification of changes in the US healthcare system which would facilitate improved care for patients	Hiss & Greenfield, 1996
Development of quality of life measures	Lomas *et al,* 1987
Appropriateness of junior doctors work out-of-hours	McKee & Black, 1993
Appropriate indications for:	
– coronary angiography	Bernstein *et al,* 1992
– abdominal aortic aneurysm surgery	Ballard *et al,* 1992
– percutaneous transluminal coronary angioplasty	Hilborne *et al,* 1992
– prostatectomy	Hunter *et al,* 1994
– total hip replacement	Imamura *et al,* 1997
– coronary artery bypass grafting	Leape *et al,* 1992b
– cataract surgery	Lee *et al,* 1993
– carotid endarterectomy	Matcher *et al,* 1992
– cholecystectomy	Scott & Black, 1991b
– spinal manipulation for low back pain	Shekelle *et al,* 1991
– hospitalisation of adolescents with conduct disorder and/or substance abuse	Strauss *et al,* 1995

The RAND form of an NGT

A nine-member group of experts is convened. These experts first define a set of indications to reflect their concepts of the critical factors (or cues) in decision-making for patients with the condition. The participants are chosen because of their clinical expertise, influence, and geographical location. Furthermore, they may represent academic and community practice and different specialties.

After agreeing on definitions and the structure of the indications (scenarios), the participants rate the indications using a 9-point scale in which 1 = extremely inappropriate (risks greatly exceed benefits), 5 = uncertain (benefits and risks about equal), and 9 = extremely appropriate (benefits greatly exceed risks). By 'appropriate', it is meant that the expected health benefits to an average patient exceed the expected health risks by a sufficiently wide margin to make the intervention worthwhile and that the intervention is superior to alternatives (including no intervention).

The final ratings of appropriateness are the result of a two-stage process. The indications are initially rated independently by each participant without discussion or contact with the others. The group then assemble and the collated ratings are presented for discussion. After discussion, each participant independently and confidentially re-rates each indication. The median rating is used as the appropriateness score.

To determine agreement and disagreement a statistical definition using the binomial distribution is applied. For a nine-member group, agreement exists when no more than two individuals rate a particular indication outside a 3-point range (i.e. 1–3, 4–6, 7–9). Disagreement about an indication exists when three or more rate a particular indication 7–9 and another three rate the same indication in the 1–3 range. Other indications are regarded either as equivocal (agreement at the centre of the scale) or as partial agreement.

Based on Bernstein et al, *1992*.

about presenting those ideas and, by separating the idea generation and discussion phases, more ideas will be developed and discussed.

The NGT attempts to structure the interaction that follows by means of a facilitator. Each idea is discussed in turn. Thus, all ideas will be discussed, rather than focusing discussion on only one or two ideas. Controlling the interaction so that all participants have the opportunity to express their views is said to reduce the dominance of the discussion by one or two vocal members.

Consensus development conference

The US National Institutes of Health have run more than 100 conferences on a variety of topics (Ferguson, 1996). Their formal guidelines for running these conferences have been modified over time. The method has subsequently been used in other countries, including Canada, the UK and Sweden (*Table 4*).

A selected group (of about ten people) is brought together to reach consensus about an issue. The format involves the participants in an open meeting, possibly over the course of a few days. Evidence is presented by various interest groups or experts who are not members of the decision-making group. The latter then retire to consider the questions in the light of the evidence presented and attempt to reach consensus. Both the open part of the conference and the private group discussion are chaired.

The development of consensus conferences has drawn on aspects of judicial decision-making, scientific conferences and the town hall meeting (Lomas, 1991). Like a legal trial, the group (jury) hear evidence on which they will later deliberate; unlike a trial, the group members are allowed to ask questions, their chairperson is responsible for controlling the proceedings, and the audience (members of the public) can also participate in the discussion. The group discussions follow an informal format (similar to a jury) with the chairperson directing discussion and delegating tasks. Although the group is encouraged to attempt to reach consensus, members are also encouraged to include minority or alternative views where consensus cannot be achieved. Although the consensus conference was developed from a need to make decisions in a public forum, rather than as a response to research on group decision-making techniques, it has mostly been evaluated in terms of its decision-making properties.

Other methods

For completeness, a number of other formal methods for aggregating the decisions of a number of people should be mentioned, though they have not commonly been used in the health sector. They include the following.

Staticised groups (also termed nominal groups, but not to be confused with the NGT) are collections of individuals who work on a problem or issue independently with no interaction. Their views are

TABLE 4 *Examples of consensus development conferences in the health field*

Subject	Reference
Treatment of stroke	*BMJ,* 1988
Arrhythmias	*Can J Cardiol,* 1994
Coronary thrombolysis	*Can J Cardiol,* 1995
Lyme disease	*Can Med Assoc J,* 1991
Assessing dementia	*Can Med Assoc J,* 1991
Treatment of early stage breast cancer	*Conn Med,* 1991
Overweight, obesity and health	Crepaldi *et al,* 1991
Venous thromboembolism	Haas, 1993
Treatment of acute otitis media	Karma *et al,* 1987
Therapeutic strategies for schizophrenic psychoses	Kovess, 1995
Management of hypertension	*Med J Aust,* 1994
Surgery for epilepsy	NIH CDP, 1990
Diagnosis and treatment of early melanoma	NIH CDP, 1992a
Diagnosis and treatment of depression in late life	NIH CDP, 1992b
Impotence	NIH CDP, 1993
Optimal calcium intake	NIH CDP, 1994
Total hip replacement	NIH CDP, 1995a
Ovarian cancer	NIH CDP, 1995b
Coronary artery bypass grafting	Stocking, 1985
Progestagen use in post-menopausal women	Whitehead & Lobo, 1988
CDP = Consensus Development Panel	

aggregated statistically and the result is treated as a group view. Research suggests that for types of problems requiring little depth of analysis, staticised groups tend to outperform methods involving interaction, but that for more complex tasks requiring a deeper analysis, interaction is beneficial (Rohrbaugh, 1979; Steiner, 1972).

Social judgement analysis (SJA) is derived from social judgement theory (Hammond & Brehmer, 1973; Rohrbaugh, 1979) and focuses on the type of feedback given to participants. Social judgement theory suggests that differences between individuals' judgements occur because of differences in the importance they attach to information and how they relate the information to their judgement. SJA attempts to map the underlying structure of an individual's decision and to provide this information as 'cognitive feedback' to the participants, so that the focus of discussion is on the logic (or lack of logic) behind the judgements. Although private decisions must be elicited to generate the individual's judgement model, the form of interaction is of secondary concern and may or may not be structured. This method is essentially a form of feedback, rather than a comprehensive consensus method. It may be valuable when seeking to understand why there is a lack of consensus on a topic, and thus may be useful for looking at variations in ratings of the appropriateness of a healthcare intervention.

Other methods focus on structuring the interaction among group members. Some methods use either a facilitator or instructions to participants to ensure that discussion passes through a series of problem-solving steps in a systematic manner – analyse the problem, generate alternative solutions, evaluate alternative solutions (Jarboe, 1988). Other methods give detailed instructions to group members on the best way to proceed, for example, avoid arguing for your own ranking and avoid changing your mind simply to avoid conflict (Hall & Watson, 1970).

Apart from these methods, there are also numerous variations on free interaction. These concern:

- the type of leadership, such as problem centred leadership (PCL; Miner, 1979)
- the group process, such as the 'step-ladder technique' in which individuals' views are added to the group one by one and discussed in turn (Rogelberg *et al,* 1992), or 'snow-balling' in which individual opinions are gathered by means of increasingly large groups
- how the decision is made, such as selecting the 'best member' (Sneizek, 1990).

Summary

In this chapter we have explained why consensus development methods are needed for producing clinical guidelines, suggested why formal methods are superior to informal or unstructured approaches, and outlined the principal features of the most commonly used methods in the health field (both for clinical guideline development and for other purposes).

Chapter 2

How do individuals and groups make decisions?

Introduction

Consensus decision-making is a complex process which involves decisions at both the individual and group level. Most of what we know about individual and group decision-making is based on extensive work in psychology. The purpose of this chapter is to highlight areas of psychological research that are central to consensus decision-making.

Decision-making by a group of individuals involves a number of psychological processes which have been the focus of extensive investigation, particularly in cognitive and social psychology. For example, consensus judgements require that people attend to, process and recall new information presented through written materials and discussion with other group members. Evaluating such information involves integrating numerous facts and opinions, weighing the relevance and strength of this information in the light of the specific judgement task, estimating probabilities for different outcomes and determining the value of alternative decisions. In consensus methods which involve face-to-face interaction between participants, these decisions are made in a social context.

The major contribution of psychological research on decision-making has been to illuminate the myriad of biases which operate at the intra-individual, inter-individual and intergroup levels when people process information to make decisions. Research in cognitive and social psychology has challenged the assumption that people process information in a logical, rational way by highlighting the irrational aspects of human judgement. Cognitive psychology addresses the biases involved in processing factual information and in making decisions; social psychology addresses the biases involved in processing social information, provided through interactions with the leader and other group members, and in making decisions in a social context (Heider, 1958; Kelley, 1967; Nisbett & Ross, 1980; Nisbett & Wilson, 1977; Weiner, 1974).

There are two types of biases that influence the way in which people process information in order to make decisions: cognitive biases and motivational biases. Cognitive biases arise from people's need to have a coherent and logical view of the world; motivational biases arise from people's need to satisfy their own needs and motives as well as those of the social group to which they belong. Key areas of psychological research relevant to information processing and decision-making in groups, which illustrate how and when such biases operate, are outlined below.

Attention and memory

Research in this area has examined what information people pay attention to and how that information is stored (encoded) and recalled from memory. Knowledge of what type of information is attended to, stored and recalled would be an important consideration in preparing and presenting information for use in consensus decision-making. Relevant areas of research include the following.

Attention

We know that **attention** is selective, in the sense that only certain information is attended to. An important consideration, therefore, in the context of consensus decision-making is to ensure that the information presented captures participants' attention. Research in this area has shown, for example, that information that is novel, unusual, distinctive, vivid or extreme is more likely to capture attention. Thus, case histories, which often provide more vivid information than quantitative data, will be more likely to be attended to and may, therefore, be more persuasive. The same is true for information that is relevant to or consistent with a person's beliefs and goals insofar as this kind of information is more likely to be attended to. This is an example of **confirmatory bias**.

Memory

Information can be stored in either **short-term (working) memory**, which has limited capacity, or **long-term memory** (Baddeley, 1986). Information in short-term memory is transient unless rehearsed. Retrieving information from long-term memory

can be difficult. The literature in this area addresses practical aspects of enhancing long-term **storage (encoding)** and **recall (retrieval)** of information from memory. A relevant point for consensus groups is that because of memory limitations, external devices for memory aiding should be used. For example, flip charts in front of groups and the provision of paper and pens to participants to make notes of points. This reduces the reliance on internal memory processes.

The way in which information is stored in memory influences the recall of information during decision-making. Research in this area shows that the recall of information is influenced not only by vividness and distinctiveness, which initially determine what information is attended to, but also by **priming** or the recency and frequency of use of information in memory. For example, it has been shown that recently used information has a greater influence on judgements and opinions (Tournageau *et al*, 1989). Similarly, frequently used information has been found to be more easily accessed and, consequently, frequency of use is a powerful determinant of what information is used in decision-making.

One common view of memory is that it can be seen as an **associative network** where ideas are represented by interlinking nodes. The ease with which a person is able to recall information is directly related to the number of links with other ideas in the associative network and how frequently these links are activated. Therefore, repeated exposure to information that is well-embedded in a larger context of related ideas is more likely to be recalled and to influence decision-making.

Research on **eyewitness testimony** (Loftus, 1979) has examined the ways in which memory can be distorted, thereby decreasing the reliability of information recalled from memory. Research in this area is based on the theory that the recall of information from memory is a constructive process (Bartlett, 1932), drawing in part on valid information as well as on 'theories' or confirmatory biases about what people expect should have happened. Factors such as the questioning that occurs after an event are known to distort eyewitness testimony. The fact that people change their opinion about information may be due as well to **demand characteristics**, that is explicit and implicit cues that indicate what behaviour is expected in a situation (Orne, 1969), which may lead to people giving responses in order to please the questioner.

Problem solving, reasoning, thinking and decision-making

Research in these areas has been dominated by theories and computational models concerning how people process information in conditions of uncertainty. People are often required to make decisions on the basis of insufficient or uncertain information. Research has addressed how people do this and whether there are optimal ways of doing so. As research in this area is highly relevant to consensus decision-making, specific aspects of this work will be reviewed in detail in subsequent chapters. Here, general areas of relevance are outlined.

Normative decision-making

Normative decision models refer to how people should make decisions (Baron, 1988). In this approach, decision models are developed based on principles of rationality. Normative decision-making models include Expected Utility theory, Multi-attribute Utility theory, and Bayes theorem. In general, these models decompose the decision problem into a number of elements and then put the components back together again according to formal rules. Given that these models are based on rational principles, the output of the model should provide the optimal decision.

The most well known normative model is that of Expected Utility in which the probabilities associated with different decision options are combined with the utility (or value) of that option to produce an expected utility. The decision-maker should choose the option with the highest expected utility derived in this way.

Descriptive decision-making

Research in the area of descriptive decision-making examines how people actually make decisions. In general, people's decisions are not fully rational. Because of limited capacity in processing information and options in order to make a judgement, people are more likely to search for a good-enough answer rather than the optimal solution.

Research on heuristics, cognitive processes that provide useful shortcuts for making judgements, shows that such rules of thumb can lead to biased judgements. For example, the availability heuristic refers to people's tendency to be biased in making judgements based on events that are readily accessible in memory. The fact of having recently used a particular intervention to treat a patient with the condition being considered in a consensus group may make information related to that

treatment more accessible and thus more influential. The representativeness heuristic refers to the tendency to consider events more representative of the total population than they are. For example, surgeons who see patients who are suitable for surgical intervention, and who therefore have good surgical outcomes, may overestimate the appropriateness of surgery for groups which are not referred to them. Information about base rates or overall distribution is often ignored in favour of one's own (often distorted) beliefs about the frequency of occurrence of events.

The way in which information is presented to decision-makers can also affect judgements. For example, framing effects or the way in which information is expressed, can have a significant impact on decision-making. Similarly, people often use a known value to 'anchor' their judgement and make adjustments to that value on the basis of new information. However, often they do not adjust enough. Much research has shown that people are poor at assessing and combining probabilities (Lichtenstein *et al,* 1978).

Prescriptive decision-making

Prescriptive decision-making involves the application both of normative and descriptive theories to 'real world' decisions (Bell *et al,* 1988). This approach combines the value of normative models with a more realistic conception of the decision-maker. It may be useful to decision-makers because it provides a framework for constructing and examining a decision process in an explicit and logical way (McNeil & Pauker, 1984). The decision-maker can thus examine the way in which elements of the decision are combined and evaluate the consistency and logical coherence of the decision. An examination of the way a decision is structured may highlight inadequacies in its representation, such as a lack of information. McNeil and Pauker (1984) identified the ability to perform sensitivity analysis as an important aspect of prescriptive models, since it allows estimation of how changes in the various parameters affect the decision.

Clinical decision-making

Much work in the area of clinical decision-making has examined how clinicians make decisions with a focus on diagnosis (Elstein & Bordage, 1988). This is a particularly difficult area because many of the normative principles do not apply. Much diagnosis is based on pattern recognition and only if this fails does it move into the generation and testing of hypotheses. Cues are used to generate hypotheses and further information is then gathered to test these hypotheses. This process is subject to many forms of bias, as discussed above, such as the tendency to seek information that will confirm rather than reject hypotheses. A more systematic approach can help (Meehl, 1954) but does not always do so (Berner *et al,* 1994) given the complexity of some tasks, the need to take into account several options (some of which have a very low probability of being true), and the limited time available.

Social cognition

Many of the concepts of cognitive psychology have been applied in social psychology to explain how people process social information about individuals and groups and to understand the processes involved in social cognition (Fiske & Taylor, 1984; Higgins & Bargh, 1987; Markus & Zajonc, 1985). Many of the same biases that influence the way people process factual information, described above, also operate when people process information about each other, as in a consensus group. Relevant areas of research include the following.

As with factual knowledge, attention to social information is **selective**. Salient information or information that is novel or unusual has a disproportionate influence on judgements about others relative to information that may be more valid. Thus group members may pay more attention to a leader or other group member who is distinctive in some way.

Categorisation simplifies information processing by helping to organise social information and make it more accessible when making judgements. That is, when presented with new information about people and groups, we automatically categorise it in order to help to organise and process it more efficiently. One way of categorising information is through the use of cognitive structures such as **schemata**, **prototypes**, and **scripts**, all of which may produce biases in the processing of social information. A schema (Taylor & Crocker, 1981) is an organised set of beliefs, thoughts and attitudes derived from past experience that we use to interpret current experience. A prototype is a schema, or a set of abstract features, commonly associated with groups of people or things in particular categories, such as 'surgeon versus physician' or 'specialist versus generalist'. A stereotype is an example of a schema about members of an identifiable group which can bias judgement in a group context. A script (Abelson, 1976; Schank & Abelson, 1977) is another kind of schema for a stereotyped sequence of events, such as examining

a patient or performing a specific type of surgery, that differs between individuals due to their particular experiences in that situation, but tends to have commonalities across individuals for common situations. As with factual information, priming (the effect of prior context on the recall of social information) is known to influence the accessibility of information.

The relevant point is that although cognitive structures such as schemata and scripts help to process information more efficiently, such categories are often oversimplifications. Categorisation of information may lead participants to reject good but inconsistent evidence which may result in a judgement of the information which is biased towards their prior categorisation or schemata.

Persuasion and attitude change

Research in this area investigates how people's attitudes can be influenced and changed (Petty & Cacioppo, 1981). Research in **communication** has examined the elements involved in persuasion in order to determine the optimal way of achieving attitude change (Hovland *et al,* 1953). Aspects of persuasion that have been studied include the source (communicator), content (message), recipient of the information, and the context in which it is presented; that is, who says what to whom and in what context. Findings from work in this area have identified several factors known to enhance the persuasiveness of information and the likelihood of attitude change. The persuasiveness of information does not depend entirely on the extent to which arguments are logical and sound but also on motivational and emotional factors. Relevant results of research include the following.

An important factor influencing the success of persuasion concerns the credibility, trustworthiness and likeability of the **communicator**. Research on **impression formation** provides knowledge about the processes involved in making judgements about other people, including information about what characteristics contribute most to our judgements of others, how we combine specific characteristics to form impressions of others, and how accurate our impressions are (Schneider *et al,* 1979). Knowledge about impression formation may be important when considering how specific characteristics of either the leader or participants may affect the latters' views of, for example, the credibility of the leader or of other group members. Information from a communicator who is viewed as being similar to or being from the same reference group as the recipient is more effective. Similarly, the persuasiveness of information is known to depend on the recipient's perception of the credibility or trustworthiness of the communicator, which may be dependent on social or professional status. Research in this area shows that people tend to form highly consistent impressions, even on the basis of little information.

Although judgements are determined mainly on the basis of verbal communication, non-verbal communication (Mehrabian, 1972), including information that is presented through visible (e.g. facial expression) and paralinguistic (e.g. voice quality) channels, is another important determinant of the persuasiveness of a communication.

A related area of research has examined mechanisms of **resistance to persuasion**. When people are aware they will be exposed to views to attempt to persuade them to change their attitude on a particular matter, they often become resistant to persuasion by virtue of having had time to consider their own and alternative positions (Petty & Cacioppo, 1977). Research in **social judgement theory** (Sherif & Hovland, 1961) has found that the amount a person changes their attitude about a specific matter depends on how discrepant the persuasive message is from the recipient's current attitude. For example, information is more likely to be effective when it advocates a position that is neither too close nor too far away from the person's initial position. The more positions a person finds unacceptable, the narrower the range of persuasive communications he or she will accept. Similarly, the degree of prior commitment to an opinion or idea has been shown to be a powerful determinant of persuasion; high commitment is associated with reduced persuasion. When attempts are made to change deeply held attitudes on a particular matter, people may feel threatened and react by becoming even more extreme in their initial opinion, so-called group polarisation (Brehm & Brehm, 1981).

Behaviour within groups

There are several areas of research in social psychology which have examined the myriad of factors that influence behaviour within groups. Some of this work, which is particularly relevant to consensus decision-making, is reviewed in detail in subsequent chapters.

Group composition and structure

Studies of group composition and structure have examined how differences in group size,

communication networks, leadership, and roles and norms affect individual and group decision-making. The optimal **size and composition** of groups has been studied in the practical context of decision-making groups, such as juries (Hastie *et al,* 1983), and is discussed in chapter 5. Research on **communication networks** within groups has examined the effect of various communication styles on group performance and satisfaction. For example, comparisons have been made between groups which allow open communication among participants and groups with more restricted channels of communication in which one person receives and passes on information to other members who have no direct communication with each other. Centralised networks can be efficient for simple tasks but are less efficient for complex tasks. Studies on **leadership** (Bales, 1970; Burke, 1971) have investigated the factors related to successful leadership, leadership styles, and group members' expectations of a leader. The effectiveness of leadership styles has been found to vary depending on the task.

Social influence

Research on social influence examines the ways in which people influence each other's behaviour in groups. Studies on **conformity** (Asch, 1956; Deutsch & Gerard, 1955), or yielding to group pressure when no direct request to comply has been made, and research from the related area of **social impact theory** which examines the influence of the presence of others on an individual's behaviour in a group context, are clearly relevant to decision-making in groups. Group judgements have a strong influence on individual decision-making in terms of both informational influence (based on facts) and normative influence (based on social pressure). Nonconformity is known to be influenced by a number of factors including group size, characteristics of group members, and the type of task. For example, pressure for conformity is stronger when it comes from a person's own group rather than from people seen as being members of another group (Abrams & Hogg, 1990).

Compliance

Studies of compliance (Freedman & Fraser, 1966), or behaviour that follows a direct request, have shown that compliance varies according to the type of power (coercive, reward, expert legitimate or referent) used to gain it (French & Raven, 1959; Podsakoff & Schriesheim, 1985). The effectiveness of different types of power varies with the situation. For example, coercive power is more effective when the source of power is present rather than absent.

Group interaction

Studies of group interaction have examined several of the processes described in previous sections. Group polarisation describes the finding that group judgements often tend to be more extreme than the pre-group judgements. Research in this area examines the reasons for polarisation, factors related to polarisation, and determinants of the extremity of polarisation. Studies of **minority group influence** (Moscovici, 1976; 1985) examine whether a minority can influence a majority, what characteristics of the minority are most influential (for example, consistency in their position), and whether the type of change brought about by a minority differs from that brought about by a majority.

Behaviour between groups

Behaviour between groups is also highly relevant to consensus decision-making. **Intergroup behaviour** occurs when members of one group interact either collectively or individually with another group or its members (Sherif & Sherif, 1979). Labour negotiations and international conflict resolution are examples of intergroup behaviour. This area of research may also be relevant to interactions within as well as between consensus groups because consensus groups often bring together participants from different subgroups with different characteristics such as discipline or status. Relevant areas of research include the following.

- Research on the outcomes of intergroup behaviour, such as **prejudice**, **discrimination**, **and stereotypes**.
- Research on the processes that emerge when groups interact, such as **categorisation and social identity**. Categorisation can lead people to assume similarities and/or differences where few actually exist. For example, categorisation often produces within and between group discrimination in which others are categorised as members of an ingroup ('us') or outgroup ('them') on the basis of often arbitrary criteria (Tajfel, 1982; Tajfel & Turner, 1986).
- Research on **intergroup conflict** has examined how conflict between groups can be reduced. Strategies known to work have been investigated in the literature on **cooperation** and **competition** (Deutsch, 1973) and **bargaining** and **negotiation** in groups (Pruitt, 1981; Rubin & Brown, 1975). Strategies for reducing conflict include establishing superordinate goals, ensuring equal status of groups or group members, institutional support, minimising conditions likely to foster stereotypes and maximising perceived fairness.

Group decision-making

Research on group decision-making is central to consensus decision-making. Some areas of this research are reviewed in detail in subsequent chapters. Some research has compared individual and group decision-making and shown that informal groups do not necessarily make better decisions than individuals. The question then addressed is, why might this be so?

The outcome of group decision-making is related to both **input** and **process factors** (*Figure 1*). Input factors influence the interaction which in turn affects the output (McGrath, 1984). There are three main categories of input factors related to the characteristics of (1) the individual participants, such as their skills, status and personality, (2) the group, such as its structure, size and norms, and (3) the environment, such as the nature of the task, level of environmental stress and reward structure.

Process factors include aspects such as communication within the group, exchange of information, alliances between members, and strategies for performing the task. Steiner (1972) identified input factors as the determinants of a group's potential, and generally viewed process factors as leading to losses rather than gains in potential. Such losses are thought to arise from motivational and coordination failures (Wilke & Meertons, 1994). (It is important to note that Steiner was not solely concerned with decision-making groups, but with task groups of all types, for example an assembly line.) Thus group performance can be described as:

group performance = group potential – process losses.

But process factors may also produce greater gains than would be expected from the inputs. For example, interaction may increase motivation and lead to a more effective pooling of resources which results in process gain. Process and input factors are not, however, independent but are often interrelated. Thus, for example, the particular group structure can influence the communication processes within a group. Indeed, a particular concern of applied researchers has been to manipulate the various input factors in order to minimise process loss (and maximise process gain).

Janis (1982) developed the concept of **groupthink**, the tendency for group members to seek concurrence, to explain decision-making in groups. Groupthink illustrates how group interaction can lead to poor decisions. Janis (1982) analysed a number of prominent decision fiascos (such as the US invasion of the Bay of Pigs in Cuba) and suggested that groupthink was most likely to occur in highly cohesive groups which are insulated from outsiders, have a directive leader and are engaging in a stressful decision-making task. Groups under these conditions may produce defective decisions because concerns about group unity override full and careful consideration of the options and the consequences of action.

Summary

Models of behaviour which have been generated from empirical research in social and cognitive psychology as described above are directly relevant to consensus decision-making as follows.

- Members of a consensus development group are required to draw on information both

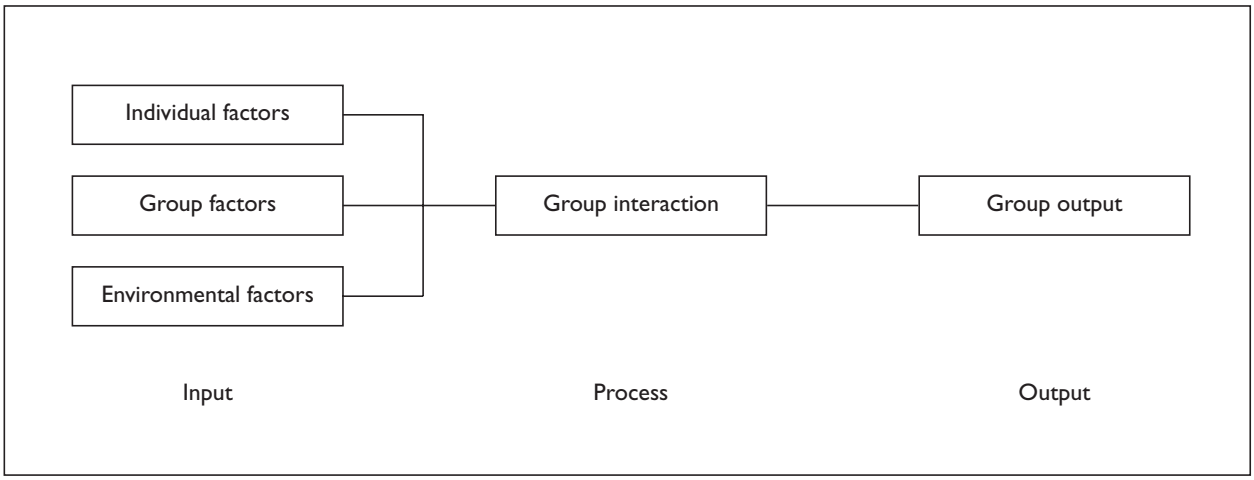


FIGURE 1 *Model of input, process and output factors influencing decision-making by groups*

from their own experience and from new information presented within the group in order to make decisions.
- The ways in which people attend to, organise and remember information affect what information is likely to be used, the impact that it will have, and the possible biases that may operate in decision-making.
- Attempts to reach consensus will involve the need for some people to change their positions, a process involving persuasion and social influences.
- Behaviour within and between group members may be influenced by the perceptions of the groups involved: do members of consensus groups see themselves as members of a common group with a common goal, or are subgroup identities more salient, perhaps leading to a conflict of interest between consensus subgroup members?

In this chapter we have explained the cognitive and social psychological theories and the work from decision analysis that underpin consensus development. The next chapter describes the conceptual model we created for structuring our review and the approach we followed for searching and synthesising the literature.

Chapter 3

Development of a conceptual model

Introduction

It is essential when reviewing any topic to have a conceptual or theoretical model to structure the task. To develop a conceptual model to guide and structure this review, we explored relevant material in the health sector, the experiences of members of the review group who had used consensus methods, and drew on psychological theory and research (outlined in chapter 2). The resulting model is described below. We then go on to explain on which aspects of consensus development we chose to focus and the strategy we adopted to identify relevant literature. Finally, three fundamental questions had to be considered:

- what are the objectives of consensus development methods?
- how can 'consensus' be defined?
- how can the validity of consensus judgements be determined?

Development of a conceptual model

Consensus development methods may be considered to involve three types of activity: planning, individual judgement, and group interaction. These activities are not necessarily clearly separated or ordered.

- **Planning.** Planning involves deciding on the process through which consensus is to be developed. Decisions must be taken about the various aspects of the process including the topic in question, who is to take part, and how the exercise is to be conducted. Materials for the consensus process must be developed. There is little literature directly concerning the planning process and little in the way of objective measures of how to do it well. Some commentators suggest that, as a minimum, details of how the various design decisions were arrived at and justifications for these decisions should be provided. It may then largely be a matter of judgement as to whether these justifications are good ones, though some aspects have been subject to more objective appraisal, for example, sampling issues.
- **Individual judgement.** The focus is on what the participants do before any group interaction and what they bring with them to the interaction. The importance of this will vary depending on the particular consensus development method used: individuals may be required to read information, to make judgements, or form opinions of the output.
- **Group interaction.** The nature of the interaction between participants from the first contact onwards varies considerably depending on the consensus development method used. In the Delphi method the only interaction is in the form of written feedback of the other participants' judgements, whereas in NGTs and consensus development conferences the interaction is face-to-face.

In addition to seeing consensus development in terms of these three activities, five components can be identified: three inputs (questions, participants, information), the process (consensus development method) and the output *(Figure 1)*. As with the three activities, the components are also interrelated.

By combining these five components and the three activities, a matrix was formed *(Figure 2)*. Each cell in the matrix describes a component in a particular activity. Brief descriptions of some of the important features of each cell are included in the matrix. We explain below what is involved in each activity and how this might impact upon the process and outcome of consensus development.

Questions

Every consensus group addresses one or more questions selected by the organisers during the planning phase. Consensus methods vary in how many questions they can handle: only a few questions (four to six) tend to be considered in consensus conferences, whereas the Delphi method and NGT can address a very large number of questions (over 1000). The design issues are who selects the questions and on what basis, how the questions are constructed, and what types of biases might be introduced as a result of question selection.

With regard to individual judgement, participants interpret the questions, form an impression of the

	Planning	Individual judgement	Group interaction
Question(s)	Selection of topic Selection of cues Comprehensiveness	Influence of cues Question structure Level of detail	Modification of question(s)
Participants	Number Type Degree of heterogeneity Selection of individuals	Representation of others Representation of self	Combination of backgrounds
Information provided for participants	Amount Selection Presentation	Read Understand Interpret	Use of information New information Feedback of group view
Method of structuring interaction	Choice of method Particular brief	Perceptions of process Past experience	Setting Structure of interaction
Output: method of synthesising individual judgements	Type Target audience Aggregation rules	Perceptions of output Acceptance	Production of output

FIGURE 2 *Matrix representation of the conceptual framework of the review. Shaded cells are those areas on which the review was concentrated*

task and, depending on the requirements of the method, express their judgements in particular formats. Concerns include how the individual makes judgements and the influence of particular types of questions and response formats on those judgements.

Interaction involves a confluence of prior individual judgements. Concerns include the role of interaction in changing/refining the question(s).

Participants

An important element of consensus decision-making is the choice of participants. As part of planning, those who are to form the group must be selected. How many participants should there be? Does it matter whether groups consist of members of just one profession or are mixed? Does it matter which particular individuals are selected? Should different participants take different roles in relation to individual judgement and interaction?

Information

Planning issues include whether to provide information, what information to provide, in what form at what stage in the process, and how to select it. Individuals will form their judgements on the basis of what they have read and understood, how they have interpreted the information they are presented with, and their assessments of its accuracy and appropriateness.

During group interaction the concern is with how information is used by the group. This includes both the use of the information received beforehand and any new information introduced during the group session.

Method of structuring the interaction

Planning decisions include what type of consensus method to use and where and how it is to be run. Is feedback important? Should it be anonymised? Should there be an audience of non-participants? Are face-to-face meetings better than more controlled interaction? Do participants stick to instructions? Is the method of combining individual judgements important? What effect does the chairperson have on decision-making?

Output: method of synthesising individual judgements

During planning, decisions must be made as to the type of output to be produced, whom the output is to be aimed at, and who is to produce it. What effect do different methods of aggregating individual judgements have on the decision? The participants may have opinions about the form

of output to be produced and interaction may involve discussing and deciding upon this.

Our approach to the review

This section describes the selection of methodological issues which formed the focus of the review, the scope of material reviewed, and the search strategy adopted.

Selection of cells

It was not possible in the given time to review in detail all 15 cells identified in the matrix (*Figure 2)*. Therefore, we chose to focus on some aspects and to mention others only briefly. The decision about the methodological issues to be focused on was based on three criteria:

- the importance of the particular aspect to consensus decision-making in the health sector
- the amount and quality of the literature available on the particular aspect
- the potential for offering practical guidance to those conducting consensus development groups.

It was not necessary for all three criteria to be met to warrant inclusion. For example, if an aspect was considered important and of practical consequence, even though the literature was sparse, it was included. Following group discussion, members of the review group indicated to which cells they would give priority. There was unanimity on two cells: method–interaction and participants–planning. Five out of six group members agreed on two more (information–planning and output–interaction) and four out of six agreed on questions–individual and method–planning. These six cells formed the focus of the review (shaded in *Figure 2*).

Selection of material to be reviewed

The type of research available and the nature of consensus development methods determined our approach to the task. The amount and type of methodological research on consensus methods used within the health sector is very limited. For example, when consensus groups are compared only a small number of groups (often only two) are used. In addition, many basic issues concerning group decision-making have not been addressed at all.

In contrast, as has been seen in chapter 2, in social and cognitive psychology there has been extensive research on issues that underlie consensus development methods, though that research has not necessarily directly examined these methods. However, although this research is relevant, because of the nature of the studies that have been carried out (the approach, subjects, tasks) it can only throw an indirect light on consensus development within the health sector. It was necessary, therefore, to glean general ideas and findings from this wider literature rather than to pursue specific findings of particular pieces of research which may be of little relevance in the health sector.

Given the scale of the literature on such topics as leadership or group decision-making, a 'systematic' approach would have been impractical and of doubtful value. Before describing the search strategy it is, therefore, necessary to explain how we defined 'systematic'.

Defining systematic

'Systematic' in one sense simply means approaching something within the framework of some explicit system. However, bound up with the notion of a 'systematic review' is comprehensiveness or completeness and reproducibility of results (Centre for Reviews and Dissemination, 1995). There are a number of aspects of a review which can be conducted in a systematic and/or comprehensive manner: defining the problem area, deciding on the sources to be searched, deciding on inclusion criteria, and synthesising the material. Different topics will be amenable to different levels of systematicity.

Through reading and a knowledge of the field we defined the factors involved in consensus methods. This conceptualisation of the problem could be defined as the 'system' through which we approached the review. The review addresses those methodological aspects that were viewed as the most important and most likely to be of interest to users of consensus methods. The aim was not to review every study within a specific area but to provide a relatively comprehensive overview of particular methodological issues which are important when using consensus development methods. Where a substantial literature was available from the health sector, we paid little or no attention to evidence from other sectors. If, however, few or no studies had been conducted in the health sector, we sought relevant evidence from other fields.

The amount of evidence needed to answer a question depends in part on the type of question asked. Questions such as 'which method is best?' almost inevitably produce an 'on balance' or 'it

depends' answer. For example, can methods for aggregating individual decisions produce different results? Other questions of the 'can this occur?' type can be answered on the basis of one study. Here a comprehensive strategy is important to find at least one relevant and high quality study, but it is not necessary to cite all similar studies.

For some answers, a systematic and reproducible synthesis of data may well involve some mathematical integration. However this is only justifiable where the number and comparability of studies are sufficient. Because of the nature of the research included in this review (either there are very few studies or there are a number of heterogeneous studies) a narrative approach, sometimes based around tables of results, was found to be more appropriate.

Literature searched and included

The majority of the literature reviewed came from published sources. Most was identified through searches of electronic bibliographic databases, though the reference lists of retrieved articles were also used. In this section, general search strategies are described along with the type and amount of literature retrieved. Each of the six cells reviewed drew on the results of the principal general search. For some cells, further focused searches were used.

The general criteria for including a reference in the review were that it dealt with a methodological aspect of consensus decision-making and it was relevant to consensus decision-making in the health sector.

Five electronic databases were searched. Medline was searched from 1966 to 1996 using the following terms.

> Consensus Panel (concensus panel), Consensus Conference (concensus conference), Consensus Development (concensus development), Consensus (near2) group, Consensus (near2) method, Expert panel, Delphi, Nominal group

Combining these terms yielded 3249 references. Many of these could immediately be discarded for a number of reasons: they referred to a different type of consensus; they referred to 'a lack of consensus' on some issue; they referred to a 'recent consensus conference'. The vast majority of the remainder of the references reported the findings of a consensus group and did not deal with methodological issues. References not discarded at this stage formed the potentially relevant articles. Abstracts of these were examined and the final articles were selected according to their relevance to consensus decision-making (thus excluding articles on topics such as 'consensus sequences' in genetics).

PsychLIT was searched from 1974 to 1996 using the following search terms.

> Group decision-making as a subject heading, Nominal and group*, Delphi, Consensus and group*, Consensus and decision*

These searches yielded 3074 articles, some of which were entirely irrelevant and some of which were of only slight relevance. In addition, separate searches were undertaken to identify potentially relevant articles for particular aspects of consensus decision-making, including the following terms.

> Expert and decision*, Expert and judgment*, Information and decision*, Framing effect*

Again, after excluding irrelevant references, the list of potentially relevant articles was further reduced.

The Social Science Citation Index was searched from 1990 to 1996 using the following search terms and yielded 166 potentially relevant articles.

> group decision-making, consensus group, consensus decision, Delphi, Nominal group

ABI inform and Sociofile were also searched, though the relevant papers found largely duplicated those from other databases.

From the searches and reference lists of articles a total of 177 empirical and review articles were selected for review.

Some preliminary considerations

Before starting the review, three fundamental questions had to be considered.

- What are the objectives of consensus development methods?
- How can 'consensus' be defined?
- How can the validity of consensus judgements be determined?

What are the objectives of consensus development methods?

One objective is to arrive at a single statement or set of statements that all participants accept (or at least no one disagrees with strongly enough to veto

the agreement). Clearly, if participants persist in disagreeing, the consensus statement(s) will have little or no content. When the process goes through a number of stages, participants have the opportunity to revise their views in the light of discussion and new information. This allows them to identify which aspects of their position are relatively unimportant to them and so can be abandoned.

The other type of objective is to identify any 'central tendency' among the group and the degree of spread of opinion around it. Again the consensus development process may give participants the opportunity to revise their views in the light of discussion and new information.

Thus on the one hand there is an attempt to facilitate consensus and on the other hand there is an attempt to describe the level of agreement. While the first of these aims is the principal goal when drawing up clinical guidelines, the latter aim is also of interest. There is value in differentiating between areas of clinical practice in which there is close, moderate or little agreement.

How can 'consensus' be defined?

The answer depends on which of the two objectives is being addressed. When prioritisation is the objective, the output of the consensus method will typically take the form of a rank ordering of a set of alternatives. Each participant's scores or ranks are pooled to arrive at a group ranking.

The production of clinical guidelines generally involves weighing the balance of benefits and risks in order to estimate the parameter, such as the effectiveness of a treatment for a series of different categories of patient.

Technical questions (judgement needed because of insufficient data) need to be distinguished from value questions (judgement needed about competing social goals), because in general they should involve different types of participant. This is an important distinction because there can be no 'correct' answer with value questions, whereas for technical questions there is a correct, if undiscovered, answer. The distinction between these two types of question is, however, often ignored.

How can the validity of consensus judgements be determined?

How do we ensure that in general we make – or recognise – 'good' judgements? The nature of a 'good' judgement depends critically on the question being asked and, by definition, we can rarely **know** whether any particular judgement is a good one at the time. The best we can do is try to identify a **method** of arriving at judgements that will, on average, produce more good judgements than other methods, or produce fewer bad judgements than other methods.

Although we might not be able to say at the time whether a particular decision is good or bad, we might, if we can identify a method with these properties, be able to support the use of that method over alternatives. This still leaves us with the problem of how to evaluate whether a method produces good or bad results, and in turn, some way of assessing what a good or bad result is.

There are five possible ways of assessing validity: comparison with 'gold standard', predictive validity, concurrent validity, internal logic and usefulness in terms of commitment and implementation.

Comparison with 'gold standard'

Test the method on questions that have correct answers which the participants do not know with any precision, for example, using almanac questions of the type 'What is the diameter of Jupiter?'. For estimations of probability, normative probabilities may be available to act as a 'gold standard'. For ranking tasks, such as the NASA moon problem (which requires participants to rank items in terms of their value for survival on the moon), rankings can be compared with rankings by experts. However, it is important to recognise that the judgements required in these types of studies are different from those required for clinical guideline development. To say that a method is good for one kind of decision does not necessarily mean that it will be good for another type of decision. Another possibility might be to compare judgements with decisions derived from a normative approach. Bernstein and colleagues (1997) compared ratings of the appropriateness of coronary revascularisation made by a Dutch and a US panel with a decision analytic model. The Dutch panel's ratings were more similar to the decision analytic model than were the US panel's ratings (see *Table 5*).

Predictive validity

In forecasting it is possible to look at whether the forecast that was made 'came true'. A good decision is one that accurately predicts future events. In clinical guideline development, however, consensus methods are used when there

TABLE 5 *Agreement between US and Dutch groups and the decision analysis (DA) model (Bernstein* et al, *1997)*

	Proportion total agreement (%)	kappa	Agreement on inappropriate cases (%)	Agreement on appropriate cases (%)
CABG				
US panel vs. DA model	89	0.18	20	94
Dutch panel vs. DA model	96	0.79	81	97
Coronary angioplasty				
US panel vs. DA model	63	0.00	2	77
Dutch panel vs. DA model	92	0.83	89	94

is uncertainty about what is correct, though at some point in the future it might be possible to compare the results of an earlier consensus process with research-based evidence as it becomes available. Even this approach is not without problems as the consensus result may still have been the best decision at the time. In other words, is it fair to judge a decision in terms of evidence unavailable at the time the decision was made? Future evidence can only determine whether or not the decision was correct; it cannot necessarily determine whether it was the best decision.

Concurrent validity

One implication of the above is that the validity of a decision should be assessed concurrently. If a decision conflicts with research-based evidence, without good reason, we can say that it is invalid. For example, Merrick and colleagues (1987) examined the validity of judgements of the appropriateness of carotid endarterectomy produced by a nine-member mixed specialty group. For those judgements that could be compared with recommendations in the literature, they found that the pattern of ratings assigned by the group and the rank ordering of indications in the literature were nearly identical. Those indications which were uncontroversially endorsed as reasons for performing endarterectomy were rated appropriate by the group, while those with high endorsement but some counter claims were judged equivocal.

Internal logic

Instead of comparing a group decision with external evidence, an alternative concurrent approach is to look at the internal logical order of a consensus group's output. In this way, the consistency of the decisions can be determined. Merrick and colleagues (1987), for example, showed that categories of patients with a higher risk of stroke without surgery were judged more appropriate for carotid endarterectomy than those with a lower risk. They found only one slight discrepancy to this logic. Hunter and colleagues (1994) also examined the internal logic of participants' ratings of appropriate indications for prostatectomy. They devised scores derived from weights on four dimensions of indications for surgery (type of retention, symptom severity, symptom type, and co-morbidity). High scores would indicate surgery was appropriate and low scores that it was inappropriate. Panellists' ratings of the indications were generally consistent with these scores.

Usefulness in terms of commitment and implementation

In the context of clinical guideline development, a good decision-making process might be seen to be one that produces results that are not only correct but also influential. However, this is a difficult definition to defend because whether guidelines are used or not may have little to do with the quality of the decision (see appendix 1 for a brief account of the impact of clinical guidelines). This might be seen as a form of face validity.

Thus, for clinical guidelines there is no absolute means for judging at the time whether a decision is valid, and thus whether a particular method for producing consensus is valid. This leaves us with the problem of how to evaluate different consensus development methods. Our focus has been to look at the factors which are likely to influence the process of consensus development, and where possible how these factors might influence the outcome.

Organisation of the review

The review is organised around the five major components of consensus development described above (*Figure 2*): questions, participants, information, method and output. Each of the next five chapters is based on one component, with the particular cell of interest forming the focus and other cells being discussed where relevant. A summary of what is known about consensus development methods is

provided at the end of each section. Each chapter concludes with a discussion of the implications of each finding for the specific practice of clinical guideline development. Having reviewed the influence of a variety of factors on aspects of consensus decision-making, we then suggest, where possible, whether the influence is likely to be beneficial or harmful to the process. Finally, in chapter 9, recommendations on the use of consensus development methods for establishing clinical guidelines and priorities for further methodological research in this area are addressed.

Chapter 4

Setting the task or questions

Introduction

Judgement of the appropriate use of investigations and treatments may be based on a few general questions (such as, 'in what circumstances would you recommend treatment X?') or a more detailed set of sub-questions in which hundreds of possible clinical scenarios may be considered. Consensus development conferences usually ask only a few questions. For the Delphi method and the NGT, the task usually involves judgements over a wide range of scenarios.

Where the task consists of judging many scenarios, the **planning** input may seem greater than where a few broad questions are asked. However, even broad questions require considerable thought and must be clearly stated. Issues that might affect the eventual consensus judgement and that therefore need to be considered include the following:

- who selects the questions?
- how the questions are selected
- how the particular cues used in the scenarios are chosen
- how representative or comprehensive the scenarios are.

In this chapter we focus on four questions about how the construction of the task may influence **individual judgements**.

- Do the particular cues included in the question influence judgement?
- Does the way a question is posed and the level of detail provided influence judgement?
- Does the way judgements are elicited influence those judgements?
- Does the level of comprehensiveness or selectivity of the scenarios affect judgement?

During **group interaction** each individual's view of the question and his or her judgement are presented and discussed. The main issue at this stage is the value of encouraging modification of the questions as a result of group discussions, particularly if any ambiguity exists in the phrasing of the questions.

Note that it is difficult to separate completely the 'questions' and the 'information' in decision-making (see *Figure 2*) as the very content and format of the questions both provide information and indicate to the participants the relevant cues. Aspects of information which are more relevant to the provision of evidence will be considered in chapter 6.

Do the particular cues included in the question influence judgement?

The questions or scenarios presented to group members involve a combination of relevant cues, the values of which are systematically varied. For example, participants may be given information about the length of time the patient has had symptoms, his or her age, and the value of a test result. The important point is that judgements of appropriateness are based on cases described according to the cues presented, and thus it is important that these cues are the ones that are most relevant to the decision. If a relevant cue is not included, or if irrelevant cues are included, how will the judgement be affected?

No research on this issue has been carried out in the health sector. A study of car mechanics by Fischhoff and colleagues (1978) showed that the cues that are included may lead the individual to close their minds to the possibility that other, excluded, cues may be relevant to the decision. In that study experts and novices judging the causes of car malfunction were presented with fault trees (diagnostic checklists presented in branching form). With both experts and novices the apparent comprehensiveness of the presentation (though there were branches missing) led to a failure to consider other causes. In general, decision-makers often use only the information given in the problem as presented to them (Slovic *et al,* 1988).

It is possible, however, that people will in fact use cues that have not been presented **if that information can be inferred**. This has not been examined with regard to formal consensus development methods, but Shanteau and Nagy (1979) demonstrated its importance in a study of people's choice of an individual with whom they would like a date. Shanteau and Nagy presented the participants

with photographs and asked them to decide on the basis of a number of cues whom they would ask for a date. The perceived likelihood that the person would accept the offer was an important element in the decision, even though this was not one of the cues presented.

As well as the clinical cues of appropriateness that are included in scenarios, it is also possible that more general, unspecified cues may be influential. One of great importance is resource availability. Judgements of appropriateness based on unlimited resources are likely to differ from judgements of appropriateness in situations where resources are constrained. No studies have addressed this by varying the 'resources available' cue and comparing the decisions. However, some studies have compared the judgements of groups from different socio-economic environments in which the availability of resources for health care differ. Brook and colleagues (1988) compared groups of doctors in the USA and the UK who were asked to provide appropriateness ratings for indications for coronary angiography and coronary artery bypass grafts (CABGs) (*Table 6*). The UK group rated 46% of indications lower (less appropriate) than the US group (only 2% were rated higher in the UK). That is, the UK group judged many indications inappropriate or equivocal when the US group judged them appropriate.

Further, though less dramatic, evidence of the influence of the social context came from a comparison of appropriate indications for cholecystectomy (removal of the gall-bladder) (Fraser *et al,* 1993). A UK group judged only 13% of scenarios as appropriate for surgery compared with 22% judged appropriate by an Israeli group. In contrast, in a similar study, the proportion of appropriate indications for total hip replacement in the UK (19%) was found to be similar to that in Japan (20%) (Imamura *et al,* 1997).

In a conference abstract, Vader and colleagues reported a comparison between a US and a Swiss group (JP Vader *et al:* unpublished communication, 1996). Participants were provided with the same literature and a similar set of questions. They provided appropriateness ratings using what the authors called a modified Delphi method. When comparable indications were rated, 80% were assigned to identical categories in the two countries, which implies that 20% were rated differently.

More recently, Bernstein and colleagues (1997) compared a US and a Dutch group rating the appropriateness of percutaneous transluminal coronary angioplasty (PTCA) and CABG, using a modified NGT (*Table 7*). They divided ratings into inappropriate and not inappropriate (appropriate plus equivocal). Overall agreement was 64% for PTCA and 90% for CABG but the kappa

TABLE 6 *Median US and UK group ratings of indications for coronary angiography and for CABG by appropriateness category (Brook* et al, *1988)*

US rating category	CABG			Coronary angiography		
	UK category			UK category		
	Appropriate	Equivocal	Inappropriate	Appropriate	Equivocal	Inappropriate
Appropriate	78	86	41	85	69	33
Equivocal	3	63	91	4	25	42
Inappropriate	0	5	113	0	2	40
Total		480			300	

Larger numbers in the top right than in the bottom left of each part of the table (for CABG and coronary angiography) indicate lower thresholds for intervening in the USA

TABLE 7 *Agreement between US and Dutch groups using a modified NGT (Bernstein* et al, *1997)*

	Proportion total agreement (%)	kappa	Agreement on inappropriate cases (%)	Agreement on appropriate cases (%)
CABG	90	0.18	20	95
Coronary angioplasty	64	0.03	5	78

values were very low (0.18 and 0.03). The US group rated 1.1% of PTCA cases (or scenarios) inappropriate whereas the Dutch group rated 37% as inappropriate. For CABG, the US group rated 1.4% inappropriate while the Dutch group rated 11.4% inappropriate.

Most of these studies show differences in ratings between countries. Why do the US groups rate more clinical indications as appropriate? It may be that the US groups interpret the evidence differently from groups of other nationalities. That is, on purely scientific grounds, the groups differ in terms of whether they believe the evidence supports treatment or not. This seems less likely than the possibility that differences in socio-economic environments in terms of resources (and maybe risk of litigation) account for the differences found.

While it is clear from these studies that the choice of cues influences judgements, participants in consensus development methods may be unaware of how much importance they attach to each cue. This has been revealed through SJA in which ratings of scenarios were used to derive statistically the implicit judgement policy of individuals (Brehmer & Joyce, 1988). In a study of decisions to give patients blood transfusions, Brown and colleagues (1992) found that doctors agreed on the rank order of the importance of four cues, though their statistically derived judgement policies showed considerable variability. Similarly, Evans and colleagues (1995) examined the prescribing of lipid lowering agents and found discrepancies between doctors' stated policies and those revealed through analysis of their ratings of scenarios. Another example, cited in chapter 3, concerned the appropriate indications for prostate surgery (Hunter *et al,* 1994). Analysis of the group's decisions revealed the importance participants attached to the various cues (such as severity, co-morbidity, age) they were asked to take into account. The fact that participants' implicit values may differ from their expressed views is one reason why guidelines derived using consensus development methods may not be fully implemented in practice.

Summary

The particular cues included in a question addressed by consensus development groups influence individual and group judgements. The use of explicit cues in the design may lead participants to ignore or undervalue other cues that they may otherwise have included. This applies both to specific clinical cues and to general contextual cues, such as the level of resources available. Individuals may not be aware of the importance they attach to specific cues and, when they are aware, the importance they actually attach to a particular cue may differ from the importance they believe they attach to it. This may lead to difficulties in resolving differences during interaction.

Does the way a question is posed and the level of detail provided influence judgement?

Clinical guidelines tend to be focused either on the best way of managing a particular condition (such as the best way of managing gall-bladder disease) or on the best way of using a particular intervention (such as the appropriate use of cholecystectomy). In the latter case only one intervention is made salient whereas in the former a number of interventions are made salient. This raises the question of whether differences in the way a question is posed influences judgement.

The level of detail (or specificity) of the question may also affect judgement. Tversky and Koehler (1994) showed that judgements of the probability of events differed depending on whether they were described in general or in detail. For example, the probability of death due to accident was judged to be lower than the sum of the probabilities of death due to traffic accident, drowning, electrocution, or any other cause. This may be due both to reminding people of possibilities they may have overlooked and to making particular possibilities more salient.

It has been shown that estimations of the probability of different diagnoses for abdominal pain are influenced by the level of detail provided (Redelmeier *et al,* 1995). A group of doctors was asked to rate the probabilities of two specific diagnoses – 'gastroenteritis' and 'ectopic pregnancy' – and the probability that the diagnosis was 'neither'. A second group was asked to provide probability estimates for each of those three categories and for two additional diagnoses. A logical assumption would be that in the latter case the two further categories and the 'none of the above' category would equal the probability of the 'none of the above' category of the first group. In practice, in the second group, the sum of the two extra diagnostic categories and the 'none of the above' category was greater than for the 'none of the above' in the first group (69% versus 50%). In other words, the level of detail specified in the task affected not only the options participants considered but also their estimation of probabilities.

Redelmeier and colleagues (1995) also presented fourth-year medical students with a scenario and one of two sets of instructions. One set included the information that 'many diagnoses are possible given this limited information, including sinusitis [a relatively harmless, non-life-threatening condition]'. The other set contained the same preamble but included along with sinusitis four other possible diagnoses (CNS vasculitis, lupus cerebritis, intracranial opportunistic infection, and a subdural haematoma), all of which are potentially life-threatening. The students were then asked to indicate whether they would recommend a computed tomography (CT) scan of the head. Fewer respondents recommended a CT scan when sinusitis was the only possible diagnosis suggested (20% versus 32%). In other words, stimulating individuals to consider possibilities they might not otherwise have considered affected their decision.

A further difference between disease-oriented and treatment-oriented judgements may be the type of decision each requires. Deciding the appropriate treatment for a disease involves a choice of treatment options. Deciding on the appropriate use of a particular treatment involves a choice of patients.

Redelmeier and Shafir (1995) provide some evidence that judgement may be affected by the number of treatment options provided. They suggest that including more options can increase the difficulty of the task and alter the result. In one example, two groups of doctors were presented with a scenario involving a patient with osteoarthritis. When their choice was restricted to no treatment or one medication, treatment was opted for in 72% of cases. When the choice was increased to include two possible medications, treatment was recommended in only 53% of cases. The difficulty in deciding between the two medications led some doctors not to recommend medication at all.

In a second example, Redelmeier and Shafir (1995) studied two groups of neurologists and neurosurgeons who were given a scenario in which they had to select which patient was to receive carotid endarterectomy (given limited treatment facilities). Again, decisions were affected by the number of patients to choose from. Redelmeier and Shafir concluded that additional similar options can increase the difficulty of decision-making and thus doctors may opt for a distinctive option or maintenance of the *status quo*.

Summary

Although research has not been extended to decision-making by consensus development groups considering the appropriateness of healthcare interventions, existing research does suggest that the way a question is posed may influence judgements about appropriateness.

The level of detail specified in a task affects not only the options participants may consider but also the participants' estimation of probabilities. Although research has not been conducted on this issue within consensus judgements of appropriateness, it is possible that judgements derived from general questions may differ from those derived from specific questions.

Differences may occur between appropriateness ratings derived when the starting point is a type of patient and those derived when the starting point is an intervention.

Does the way judgements are elicited influence those judgements?

When participants are asked to provide a global judgement of the appropriateness of an intervention, their decision will be influenced by implicit estimates of the probabilities and values of different outcomes. In general the appropriateness of a treatment will reflect the balance of probabilities between achieving a positively valued outcome (such as survival) and a negative outcome (such as death). Usually the probabilities and values associated with outcomes are not brought out explicitly, as only a simple global judgement is required.

One study of consensus development methods which has examined this issue compared global judgements of appropriateness with judgements derived from decision analysis based on the same individuals' estimates of probabilities and utilities (Oddone *et al,* 1994). Using a 9-point appropriateness scale, group members each rated the appropriateness of 17 scenarios for carotid endarterectomy. After discussion they re-rated the items. The scenarios were then assessed by means of a decision model which incorporated the group members' estimates of probabilities and utilities for clinical variables which influenced the decision to perform carotid endarterectomy. From these estimates the expected utility, expressed in quality-adjusted life-years (QALYs), was derived.

The correlation between the median global rating and the median number of QALYs was high (Spearman's rank correlation coefficient

0.88). The median number of QALYs for indications globally rated as appropriate was 0.69, for those rated equivocal the median number of QALYs was 0.37 and for those rated inappropriate the median number of QALYs was –0.04. Oddone and colleagues (1994) concluded that global judgements were consistent with the probability and utility estimates which implicitly underlie global judgements. However, the expected utility of carotid endarterectomy was generally higher than global estimates. These findings suggest that judgements may be influenced by the form in which those judgements are elicited (global assessment or estimates of component probabilities). Any influence, however, is slight given the close association between the two.

McClellan and Brook (1992) also compared global ratings of the appropriateness of carotid endarterectomy with probability estimates from the same doctors. Participants rated six probabilities: 30-day mortality with or without surgery, and 30-day and 1-year stroke rates with or without surgery. From this McClellan and Brook generated a 1-year healthy outcome rate. Correlations between doctors' global ratings and probability estimates were poor. Only two of the eight group members had significant associations between the two measures.

In a similar study, Silverstein and Ballard (1998) asked a nine-member group to rate the appropriateness of elective resection for abdominal aortic aneurysms (AAAs). The group also estimated the probabilities of death (within 30 days, 1 year or 5 years) from AAA-related causes and from unrelated causes both for patients who underwent surgery and for those who did not. Surgery was defined as appropriate if there was a 5% or greater increase in the probability of 5-year survival, equivocal if the increase was between 0–5%, and inappropriate if there was a decrease in 5-year survival. These probability estimates differed significantly from the group's global ratings of appropriateness: more indications were rated appropriate according to probability estimates than according to global estimates (49% and 36%, respectively). Across all participants there was poor agreement between the two methods (kappa 0.28). Poor concordance (kappa < 0.40) was noted for two-thirds of the participants. Silverstein and Ballard (1998) suggest that global judgements may not be simply based on probabilities but also on the values or utilities of the outcomes, as was found in the study by Oddone and colleagues (1994).

Summary

The answers obtained from a group may be influenced by the way in which the judgement is elicited. Global views on appropriateness do not necessarily reflect participants' probability estimates of different outcomes. This is because global views also take values or utilities of different outcomes into account. Little work on this has been reported within the field of consensus decision-making.

Does the level of comprehensiveness or selectivity of the scenarios affect judgement?

One reason for low agreement among participants in ratings of appropriateness based on scenarios may be due to the inclusion of all possible scenarios, regardless of how rarely a particular scenario occurs. Rare scenarios are included so as to be exhaustive of all possible indications. Thus many of the scenarios may never or only rarely occur in practice and are unfamiliar even to specialist doctors.

Park and colleagues (1989) examined this possibility by comparing ratings of theoretical indications with indications reflecting actual clinical practice for three procedures (coronary angiography, endoscopy and carotid endarterectomy). Using a modified NGT, nine-member mixed specialty groups rated lists of possible indications. To identify cases for which the procedures were used in practice, the authors randomly sampled medical records in five geographical areas in the USA for patients aged 65 years and older and obtained a sample of approximately 1500 cases per condition. Each case was categorised according to whether or not there was agreement about the appropriateness of treatment (*Table 8*). Only 20–40% of the theoretical indications actually arose in practice and a much smaller proportion accounted for the majority of cases: for angiography 2.7% of the 300 theoretical indications accounted for over half of the cases. For two of the procedures, cases were more likely to occur in categories for which the group had reached agreement, though the opposite was true for carotid endarterectomy.

Summary

Because groups may experience greater agreement and less disagreement when they are restricted to considering scenarios with which they are familiar, the temptation to be comprehensive by including every theoretically possible scenario may be counterproductive. Also, being selective allows larger numbers of cues to be considered explicitly which should be an advantage in terms of reliability.

TABLE 8 *Proportions of theoretical indications (TI) and of cases about which a consensus development group agreed or disagreed as to the appropriateness of treatment (Park* et al, *1989)*

	Coronary angiography		Endoscopy		Carotid endarterectomy	
	TI (%)	Cases (%)	TI (%)	Cases (%)	TI (%)	Cases (%)
Agreement	28.0	29.5	25.4	39.5	40.9	16.4
Disagreement	30.0	22.6	48.5	40.9	34.0	50.4

Implications for clinical guideline development

Selection of cues

The cues included in scenarios are an important element in judgement, and therefore considerable care must be given to their selection. A two-fold strategy for selecting cues may be appropriate. In addition to reviewing the literature on the topic, clinicians involved in the consensus group should be given the opportunity to say which cues they consider important. Having clinicians develop cues may perform two functions: it will help maintain their compliance and participation and it may help them justify their judgements if the cues they believe are important have been included. Participants' views of the relevant cues might be obtained during the first round of consensus development. However, although clinicians are likely to be aware of most of the important cues, they may not give as much importance to some cues which the literature suggests are relevant. It is important therefore also to develop cues from the literature.

Contextual cues (such as whether judgements should assume unlimited healthcare resources or the reality of restricted resources) are as important as specific cues. Clinicians are likely to make some assumptions about these cues if they are not specified in the task. It is therefore important that these cues are explicitly included in the task.

Specifying the questions

Decisions must be made about whether to focus on ways of managing a specific condition or ways of using an intervention. If the focus is on an intervention, care needs to be taken as to how other relevant interventions are dealt with because the appropriateness of any intervention is affected by whether there are other interventions which may be more appropriate. In posing questions of appropriateness, other interventions can be made more or less salient. For example, there could simply be some general statement at the beginning of the questionnaire which instructs participants to make their judgements in terms of the intervention under consideration being more or less appropriate than other interventions in general. On the other hand, specific alternative interventions could be identified. It is also possible to give each rating of the appropriateness of an intervention in the context of other interventions by asking how appropriate the intervention is in comparison with other specific interventions.

The wording of questions or scenarios needs careful consideration. Their level of detail (general versus specific) needs careful consideration because it may affect participants' judgements.

Elicitation of the judgement

The way in which judgements are elicited also needs consideration. Is a global judgement elicited, or is some attempt made to break the judgement down into probability and utility estimates? The latter is likely to be a more difficult task for participants, and because of the lack of research in this area it is unclear whether there are benefits in terms of improved judgements.

Selecting scenarios

Although including all possible scenarios may seem to increase the comprehensiveness of the exercise, if many of the scenarios never occur in practice the increased burden on the respondents may not be justified by the limited value of the information provided. It may be that judgements of scenarios which never or rarely occur in practice are less reliable than judgements of scenarios which more commonly occur in practice, though there is no research on this issue. Furthermore, requiring participants to judge what they might see as numerous irrelevant scenarios may simply alienate them from the task.

Chapter 5
Participants

Introduction

There are essentially two stages to decisions about whom to include as participants in a consensus development group. The first involves questions about the type of participant and the composition of the group. Most writers suggest that consensus development groups should be composed of people who are expert in the appropriate area and who have credibility with the target audience (Fink *et al,* 1984; Jones & Hunter, 1995; Lomas, 1991). This raises the question of what constitutes an expert. Clinicians have clinical expertise, researchers have scientific expertise, and lay people or patients have expertise from having experienced the impact of the condition or intervention. Representatives from all of these 'expert groups' may be required. Once the composition of the group has been decided, questions about the procedures for selecting, or sampling, individuals need to be addressed.

Five inter-related questions therefore have been considered.

- To what extent is a group decision affected by the particular individuals who participate?
- What effect does heterogeneity in group composition have on group judgement?
- Which personal characteristics are important influences on group decisions?
- Do different categories of participants produce different results?
- Does the number of participants matter?

To what extent is a group decision affected by the particular individuals who participate?

Some of the underlying principles of sampling theory are relevant to the selection of participants. For example, the choice of hospital or geographical area from which participants are selected might be expected to affect the outcome. Also, refusal to take part may introduce bias. Only one study in the health field has examined the characteristics of non-respondents with regard to selection for a consensus development group (McKee *et al,* 1991). It was found that agreement to participate by doctors was unrelated to years since qualification, specialty, sex, country of graduation, or possession of higher degrees.

Several studies have assessed the extent to which the outcome of a consensus development method is affected by the particular individuals chosen by comparing similarly composed groups (*Table 9*). Kastein and colleagues (1993) used the Delphi method to compare two similarly composed groups made up of family doctors and medical specialists. The groups developed evaluation criteria for the performance of family doctors consulted by patients with abdominal pain and constipation. Over two rounds the doctors replied 'yes' or 'no' to possible performance criteria. Agreement was defined as 75% in one category without strong contrary arguments. Only those items on which there was disagreement were fed back in the second round. After the final round each group had left out nine criteria developed by the other

TABLE 9 *Studies that compared similar types of groups*

Study	Reference
Physicians: appropriate indications for coronary angiography	Chassin, 1989
Nurse managers: competencies	Duffield, 1993
Mixed (GPs and specialists): abdominal pain + constipation	Kastein *et al,* 1993
Mixed (doctors): women's health issues	Brown & Redman, 1995
GPs: sinusitis, dyspepsia	Pearson *et al,* 1995
Mixed (doctors): breast cancer	Penna *et al,* unpublished communication, 1997

group, 12% of all priorities. Thus there was 88% commonality in each case.

Duffield (1993) used the Delphi method to compare two groups asked to define the competencies expected of first-line nurse managers. Respondents rated a list of 168 competencies on a 5-point scale in terms of the extent to which they agreed that the items were necessary skills. Mean scores of 3 or more defined a necessary skill. Consensus was defined as the point at which 10% or less of competencies moved from above to below a score of 3 (or vice versa). According to this definition there was high agreement between the two groups (92.9% of the competencies). However, this point was reached after only two rounds because there was substantial agreement before feedback (only four to six items moved after the first round).

Chassin (1989) reports on comparisons between two groups' ratings of the appropriateness of coronary angiography. A modified NGT was used and there were 4 years between the ratings by the first and second groups. There was a good deal of similarity in the ratings by the two groups (*Table 10*). All changes were shifts from adjacent categories (such as equivocal to appropriate) and the changes occurred in categories for which changes had occurred in cardiology since the first group's ratings.

TABLE 10 *Ratings of appropriateness by two similar groups (number of indications) (Chassin, 1989)*

	1988 group		
1984 group	**Appropriate**	**Equivocal**	**Inappropriate**
Appropriate	59	6	0
Equivocal	2	4	4
Inappropriate	0	2	19

Pearson and colleagues (1995) compared three groups of primary care doctors. They used an NGT and a modified Delphi method, though the latter, coming after the NGT, did not alter the results. The participants' task was to develop algorithms for the management of two common clinical problems, dyspepsia and acute sinusitis. All were provided with the same literature review and an initial 'seed' algorithm from which to work. The groups were compared for the clinical logic of the final algorithms they produced on a scale from 0 (different) to 10 (identical) by rating patient vignettes. Scores greater than 4 indicated that half of the vignettes were judged to have similar or identical management on the algorithms. For the dyspepsia algorithms the three groups produced similar algorithms, with a score of 6.1 suggesting that a high level of reproducibility is possible. In contrast, two groups produced similar algorithms for sinusitis with a score of 4.9, while one group produced an algorithm quite different from the others, with a score of 1.9. In contrast to the earlier studies, these differences suggest that group composition will affect the guidelines produced, depending on the question addressed.

Brown and Redman (1995) compared two groups, each with 27 members drawn from diverse specialties, using an NGT. Each group generated a list of high priority women's health issues and then ranked these in terms of their priorities as targets for health promotion. The priorities selected were similar though the rankings differed: 'healthy weight' was ranked highest by one group (22% of votes) but was near the bottom for the other group (5% of votes).

Recently, Penna and colleagues reported in a conference abstract a comparison of four multidisciplinary groups composed of doctors dealing with breast cancer (A Penna *et al:* unpublished communication, 1997). They used a modified NGT to rate the appropriateness of indications and found 4–16% 'true' disagreement between groups. More detailed information on this study has not yet been published.

Summary

Little is known about the representativeness of participants in consensus development groups. Knowledge of the impact of individuals is also limited by a lack of research. Studies that have been performed suggest the selection of individuals has some, though not a great deal, of influence on outcome. Some studies show similarity between similarly composed panels, but others show differences. In most studies the number of groups compared is small and thus the findings weak. The particular tasks and procedures used also have some shortcomings.

The Delphi method has received some attention but very few studies have compared similarly composed groups using an NGT. Thus it is not possible to make any definitive statement about whether similar groups will produce similar results. The particular application of the method is likely to be an important factor. We do not expect that, as a general class, 'questionnaires' will be equally reliable for all samples of respondents. Rather, we assess the reliability of a particular questionnaire.

So too, with consensus instruments: their reliability across samples seems highly dependent on the particular application of the method.

What effect does heterogeneity in group composition have on group judgement?

Beyond the context of health care, there is a large amount of research within social and organisational psychology on the composition of groups. A variety of characteristics have been studied including:

- demographic characteristics, such as age, sex, cultural background, and occupation
- the abilities, expertise and status of group members
- the mix of initial opinions in the group.

There have also been studies of the effect of the personality of participants but since information about this is not usually available when selecting participants its effect will not be considered further.

Some research in organisational psychology suggests that diversity in a decision-making team can lead to better performance. This is assumed to occur because diversity allows for the consideration of different perspectives and a wider variety of alternatives. Bantel (1993a; 1993b) examined the strategic clarity of banks (associated with better performance) in relation to the demographic characteristics of the top management team. She found that teams that were heterogeneous in terms of education and functional expertise had greater strategic clarity than those with more homogeneous teams. She suggested that this diversity provided a variety of perspectives which led to more thorough and creative strategic decision-making (or fewer assumptions about shared values).

Other studies have also found a positive relationship between diversity and performance (Murray, 1989; Wiersema & Bantel, 1992). Jackson (1992) reviewed research on demographic characteristics, personality, attitudes, skills and abilities and found fairly consistent evidence that groups heterogeneous in personal attributes outperformed homogeneous groups. In terms of abilities and skills, the evidence that heterogeneous groups outperformed homogeneous groups was more limited.

Other reviews, however, suggest that while heterogeneity can have a positive effect, it can also have a negative effect on performance (Guzzo & Dickson, 1996; Maznevski, 1994). These reviews, however, do not distinguish between different types of heterogeneity. Guzzo & Dickson (1996) concluded that on balance there was more evidence of positive than negative effects of heterogeneity. In contrast, Maznevski (1994) suggests that much research implies that heterogeneity is detrimental and those studies which show positive effects have other homogeneous factors within the group. In other words, heterogeneity exists with regard to some attributes and homogeneity with regard to others. If heterogeneity is detrimental to performance it may be because of the increased possibility of conflict within heterogeneous groups. However, conflict itself can have both positive and negative effects on performance (Schweiger *et al,* 1989; Schwenk & Cosier, 1993). Jehn (1995) has suggested that task-related conflict may be detrimental to routine tasks, but beneficial to non-routine tasks, leading in the latter case to more open discussion and critical evaluation of problems.

Summary

The weight of evidence suggests that heterogeneity in a decision-making group can lead to a better performance than homogeneity. There is, however, some evidence that heterogeneity may have an adverse effect because conflict may arise between diverse participants. The effect of heterogeneity depends to some extent on the task being undertaken.

Which personal characteristics are important influences on group decisions?

Given that the selection of individuals may have an impact on a group's decision, which personal characteristics are important influences? Most studies which have explored this have been laboratory-based and have focused on status and expertise or participants' initial positions.

Status and expertise

Do people who have expertise dominate the discussion, and are their views likely to be the ones adopted by the group? There is a tendency for people with higher status to attempt to influence the group more, generally with success (Levine & Moreland, 1990).

A study of six consensus development conferences (involving 86 group participants and 152 speakers) suggested there was differential influence based on status (Vinokur *et al,* 1985). Observations of group interaction showed that participation varied, with about one-third of members actively and

continuously participating, one-third intermittently participating and one-third participating little. The participation rate was related to the status of the participants and the relevance of their field of expertise to the question or task. Ratings by the participants (n = 68) of the contributions of different types of members showed a similar pattern: those most expert were rated as providing the greatest contribution (4.75 on a 5-point scale) and those least expert as providing the least contribution (2.89).

There may, however, be differences between the level of influence as perceived by group members and the actual level of influence. Bottger (1984) examined the effect of expertise using the NASA moon problem in which participants rank 15 items that might contribute to survival after a crash-landing on the moon. Expertise was defined by the correctness of the individual's solution to the problem. The amount of participation was a better predictor of perceived influence than expertise, whereas expertise was a better predictor of actual influence.

Kirchler and Davis (1986) examined how groups composed of members of equal status, of slightly differing status, and of widely differing status arrived at decisions in three tasks (one intellectual task with a demonstrably correct answer and two judgemental tasks). They then tried to find the model that best predicted the outcome. For the judgemental tasks, decisions by groups of equal or similar status were best predicted by the majority opinion rule. When members had unequal status, for one judgemental task the outcome was best predicted by assuming that higher status people were more likely to influence the decision. For the other judgemental task, both models described the data equally well. The intellectual task was best predicted by a truth-wins model (whoever gets the answer right has most influence regardless of status).

Initial position and mix of initial opinions

Other studies have investigated the initial positions of members and the mix of opinions and have shown how these factors can have a substantial impact on the group decision. Of course, when selecting participants their initial positions may not be known.

Often, the majority position rules. There is evidence from jury decision-making that when a two-thirds majority favours a position initially, that position is likely to be the final outcome (Davis *et al,* 1975). However, much research shows that group members with minority views are not without influence (Maass & Clark, 1984). Nemeth (1992) suggests that minority views can improve the quality of decision-making by stimulating divergent thinking. There is some evidence for this in studies involving simple tasks (such as identifying words in strings of letters), but little evidence in relation to more complex decision-making tasks.

A related area of research has examined how shifts from individual pre-group judgements to a final group opinion are related to initial positions. It has been found that group judgements are often more extreme in the direction of the individuals' pre-group judgements and that this occurs most strongly in groups which are homogeneous in their initial judgements. Research by Williams and Taormina (1993) found that groups in which all members held similar initial views became more extreme than groups in which the initial viewpoint was opposed by a minority. There is debate about whether such a shift is the result of information exchange or normative influence. There is some support for the idea that for fact-oriented tasks informational influence predominates, while in value-oriented tasks, normative influence predominates (Isenberg, 1986).

When there is an initial split in the group, with subgroups favouring different sides, groups tend to move towards each other's position (Vinokur & Burnstein, 1978a; 1978b). However, research by Whitney and Smith (1983) suggests that when two subgroups form cohesive groups, there may be polarisation and even conflict between them; that is, they may move further apart in their views. This may impede the flow of information, thus increasing normative influence. In contrast, when there is a lack of cohesion within subgroups, information exchange is greater and there is less polarisation between the subgroups.

Summary

The status of participants affects their degree of influence on the group. In groups of diverse status, those with higher status exert more influence. In more homogeneous groups, group decisions tend to reflect the majority view.

Initial opinions of participants affect the group process. If there is a majority view, this is likely to determine the final decision. If there is an initial consensus, a shift may occur in which the final decision is more extreme. If there is a split view initially, members will tend to move towards one another's views but this depends on the degree to which those with differing views form cohesive

TABLE 11 *Studies that compared groups of different composition, and sub-groups of mixed groups*

Study	Reference
Groups of different composition	
Physicians vs. patients: communication	Lomas *et al,* 1987
Mixed vs. surgeons: cholecystectomy	Scott & Black, 1991b
Mixed vs. surgeons: carotid endarterectomy	Leape *et al,* 1992c
Mixed vs. chiropractics: spinal manipulation	Coulter *et al,* 1995
Sub-groups within mixed groups	
Medical generalists vs. medical specialists vs. surgeons: coronary artery surgery; cholecystectomy; carotid endarterectomy	Brook *et al,* 1988; Park *et al,* 1986
Doctors vs. nurses: preventability of infant deaths	Zadinsky & Boettcher, 1992
Medical generalists vs. medical specialists vs. surgeons: cholecystectomy	Fraser *et al,* 1993
Performers vs. related area physicians vs. primary care physicians: AAA surgery; carotid endarterectomy; cataract surgery; coronary angiography; CABG; coronary angioplasty	Kahan *et al,* 1996
General practitioners vs. specialists: performance criteria for general practitioners	Kastein *et al,* 1996

subgroups. The more cohesive the subgroups, the less chance of achieving consensus and the more chance there may be polarisation.

Do different categories of participants produce different results?

Within the healthcare field, three studies have compared groups composed of different specialists or healthcare staff, and one has compared clinicians and patients (*Table 11*).

Scott and Black (1991a) used a modified NGT and compared two groups of doctors from different specialties. One group was composed of a mix of relevant specialists and the other was composed entirely of surgeons (*Table 12*). Group members individually rated indications for the appropriateness of cholecystectomy on a 9-point scale and then met to discuss their ratings. The distribution of initial ratings was fed back to participants at a group session. The surgical group rated more indications as appropriate than did the mixed group (29% versus 13%, respectively) and rated fewer indications as inappropriate (27% versus 50%).

Using a modified NGT, Leape and colleagues (1992a) also compared a mixed group of doctors with a surgical group in ratings of appropriateness of indications for carotid endarterectomy. They found that in comparison with the mixed panel the surgical panel rated more indications as appropriate (24% versus 14%) and fewer indications as inappropriate (61% versus 70%).

A group of doctors from a variety of specialties was compared with a group composed of chiropractic physicians, in a study examining appropriateness ratings for spinal manipulation for low back pain. Using a modified NGT, Coulter and colleagues (1995) found that the chiropractic physicians were more likely to rate indications as appropriate.

Lomas and colleagues (1987) compared two groups of people who had suffered a stroke and a group

TABLE 12 *Comparison of appropriateness of indications according to specialist and to mixed groups (Coulter* et al, *1995; Leape* et al, *1992c; Scott & Black, 1991b)*

Topic	Specialist group (%)	Mixed group (%)
Cholecystectomy		
Agreement	**61**	**67**
Appropriate	29	13
Equivocal	5	4
Inappropriate	27	50
Partial agreement	**31**	**18**
Disagreement	**8**	**15**
Carotid endarterectomy		
Agreement		
Appropriate	70	38
Equivocal	10	31
Inappropriate	19	31
Spinal manipulation		
Agreement		
Appropriate	33	9
Equivocal	22	37
Inappropriate	45	54

of clinicians (doctors, nurses, speech therapists) using an NGT to develop a list of important communication situations (situations in which a stroke victim has to be able to communicate). Qualitative comparisons of the lists showed similarities between the two patient groups (52% of situations were the same) but differences between the clinicians' group and the patients' groups (only 37% agreement with the patients' groups).

Another approach to investigating this question has involved comparisons of the ratings or outcomes of different specialties within a mixed group (*Table 11*). Park and colleagues (1986) and Brook and colleagues (1988) reported differences between specialty groups in ratings of the appropriateness of coronary artery surgery, cholecystectomy and carotid endarterectomy. Park and colleagues (1986) used three groups composed of medical generalists, medical specialists, and surgeons. They reported that medical specialists were always close to the overall mean rating whereas surgeons rated surgical procedures higher than the overall mean. Brook and colleagues reported similar findings when UK doctors considered the appropriateness of coronary artery surgery.

Fraser and colleagues (1993) reported similar findings in a mixed panel rating clinical indications for cholecystectomy. Medical specialists (gastroenterologists) were less likely to indicate surgery than surgeons. However, unlike in the study by Park and colleagues (1986), gastroenterologists were less likely to recommend surgery than medical generalists.

Comparisons were made between the ratings of different types of members from five groups who rated six procedures: AAA surgery, coronary angiography, carotid endarterectomy, cataract surgery, CABG and PTCA (Kahan *et al,* 1996). They divided participants into 'performers' (doctors who perform the procedure), 'related' (doctors concerned with related diseases) and 'primary' (primary care doctors). Each group was composed of three or four 'performers' and 'related' doctors, and two primary care doctors, with the exception of the cataract group which had four primary care doctors. Kahan and colleagues (1996) found that 'performers' rated more indications as appropriate than 'related' or primary care doctors. The exception to this was PTCA, for which surgeons in the 'related' category rated more indications appropriate than the cardiologist 'performers'. In general 'performers' rated procedures on average one point higher than primary care doctors, with 'related' doctors in between (*Table 13*).

In addition Kahan and colleagues computed a conformity score for ratings of four procedures. This conformity score was an average for each participant across all indications which showed the extent to which a participant changed his or her rating of an indication from round one to round two towards the round one group median. All categories of participants had mostly positive scores, indicating that they all moved toward the median. However, primary care doctors moved more than others ($p = 0.03$).

Other studies both within and outside the health field have reported similar findings using the Delphi method (Cannon *et al,* 1992; Hakim & Weinblatt, 1993). Kastein and colleagues (1993) found significant differences within groups

TABLE 13 *Percentage of indications rated as appropriate and conformity to group median by different types of physicians (Kahan* et al, *1996)*

	Proportion appropriate (%)			Conformity to group median		
Procedure	**Performers**	**Related**	**Primary care**	**Performers**	**Related**	**Primary care**
AAA surgery	38	37	29	–	–	–
Coronary angiography	58	45	23	–0.17	–0.01	0.02
Carotid endarterectomy	34	16	14	0.18	0.10	0.49
Cataract surgery	53	54	40	–	–	–
CABG	47	39	31	0.30	0.29	0.49
Coronary angioplasty	36	42	27	0.33	0.25	0.76

Performers = doctors who perform the procedure; Related = doctors concerned with related diseases; Primary care = primary care doctors

between specialists and family doctors who had the task of deciding upon criteria for assessing the performance of family doctors. Zadinsky and Boettcher (1992) found differences between nurses' and doctors' ratings of the preventability of infant deaths. However, Tepper and colleagues (1995) examined groups considering criteria for different payment methods for inpatient rehabilitation and found few differences between groups.

These results raise questions about the presentation of the results of consensus methods. To what extent is the average a good representation of the group judgement, when identifiable subgroups within the group differ in their judgements?

Summary

These studies, although few in number, show that differences in group composition may lead to different judgements. More specifically, members of a specialty are more likely to advocate techniques that involve their specialty. This may reflect their greater knowledge of the scientific evidence on the appropriate use of the technique or their limited perspective on alternative strategies. Even more dramatic contrasts may arise if healthcare professionals are compared with consumers of services. Whatever the explanation, these studies confirm that the composition of groups is important in determining the decision reached.

Does the number of participants matter?

When combining individual judgements, more is generally better. As the number of judges increases the reliability of a composite judgement increases. In a theoretical study which assumed errors of judgement around a 'true' value, Hogarth (1978) found that when there was close agreement among the participants but their opinions bore little relation to the 'truth', there were no gains in group validity (correlation of group mean with true value) to be had from using groups of more than about five people. When mean correlation between individual and true values was low, but mean correlation between individual values was even lower, increasing numbers up to about 20 gave quite respectable group validity. However, under most sets of assumptions, there was little advantage in terms of 'group validity' in increasing numbers much above ten. Paradoxically, at first sight there are circumstances in which addition of another participant who lowers the mean agreement of the group can increase group validity more than an extra participant with greater individual validity.

Huber and Delbecq (1972) also found that there was little difference in expected absolute error between group sizes of ten and 20. However, this was again a theoretical study with assumed distributions of error around 'correct' values, and there is very little actual empirical evidence on the effect of the number of participants on the reliability or validity of consensus processes.

In a study of the ratings of the quality of medical care, Richardson (1972) showed that reliability increased considerably with increasing number in the group over the range of one to ten participants and then began to level off. However, it required on average 16 to 28 judges to produce a composite judgement of the quality of care for a single case with a reliability of 0.95.

Increasing the size of groups that interact may not necessarily be beneficial because size will affect processes within groups. For example, as group size increases, participation may become more unequal (Shaw, 1981). However, in studies of group decision-making, effects due to size have been difficult to demonstrate (McGrath, 1978). Nagao and Davis (1980) showed that theoretically one would expect to find differences in decisions between groups of different sizes but, because the differences are slight, a large number of groups would need to be examined to detect any such differences. In practice, few effects of size have been found in studies comparing six-person and 12-person groups (Davis, 1992; McGrath, 1984).

Summary

In general, having more group members will increase the reliability of group judgement. However, where the group members interact, large groups may cause coordination problems within the group. Although it is theoretically likely that group size will affect decision-making, the effects are subtle and difficult to detect. It seems likely that below about six participants, reliability will decline quite rapidly, while above about 12, improvements in reliability will be subject to diminishing returns.

Implications for clinical guideline development

Selecting the particular participants

The selection of the particular individuals is likely to have little impact on the group decision as long as the group is of sufficient size. To enhance the credibility and widespread acceptance of the guidelines, it is probably wise that the selection reflects the full range of key characteristics of

the population from which the participants are drawn. And further, the basis for selection should be seen to be unbiased. For example, selection of clinicians on the basis of their acquaintance with the organisers will appear biased.

Composition of the group

Whether a homogeneous or heterogeneous group is best will depend in part on the purpose of the exercise. If the aim is to define common ground and maximise areas of agreement, groups should be homogeneous in composition. If, in contrast, the aim is to identify and explore areas of uncertainty, a heterogeneous group is appropriate. The level of controversy which exists on an issue needs to be considered when selecting group members. Steps to manage conflict in a constructive way may be necessary if there is considerable controversy with opposing groups.

As participants' status may affect their contribution to and influence within a group, efforts should be made to mitigate the effects of status (for example, by the use of confidential methods which promote more equal participation of members).

In judgements of clinical appropriateness, the most relevant background factor is medical speciality. The research demonstrates that the specialty background of participants can have a substantial effect on judgement. The homogeneity or heterogeneity of groups with respect to specialty background is an important consideration. The initial opinions of members may also be related to their specialty and thus the dynamics of groups, in terms of their initial positions, may be related to the specialty of the participants. For example, a homogeneous group of one specialty may be in general agreement at the outset in terms of their judgements of appropriateness. In this case the group may become more extreme after discussion. On the other hand, when participants represent different disciplines, and there are initially different views within the group, there may be either a moving together, or a moving apart, depending on the cohesiveness within the subgroups. Consensus-based guidelines should therefore be interpreted in the context of the specialty composition of the group.

Group size

It seems likely that with groups of fewer than about six participants, reliability will decline quite rapidly, whereas with groups of above about 12 members, improvements in reliability will be subject to diminishing returns.

Chapter 6
Information

Introduction

It is now widely accepted that clinical guidelines should be based, when possible, on the results of scientific research. Typically these results will be of the form 'if you treat patients of type A with intervention B, the expected benefits, risks and costs will be C'. It is also recognised that, in practice, guidelines have to be interpreted in the light of local circumstances and in response to atypical patients. This interpretation will draw heavily on clinical experience.

However, clinical guidelines are essentially recommendations for action of the form 'patients of type A ought to be treated with intervention B'. What is not so widely recognised is that any statement about what **ought** to be done must include some element of value judgement. Thus clinical guidelines necessarily combine description (statements about how nature works, ideally empirically based, but involving judgements because the research base is inadequate) with proscription (involving judgements about what is important).

Participants in consensus development processes may be recruited on the basis that they start with a good knowledge of the research results. Additionally or alternatively, they may contribute relevant experience. At least they will be expected to have sufficient technical background to be able to interpret any additional information, such as a review of research results, that they may be given, and to interact usefully. Thus the scientific aspect of information is generally well provided for. In general, however, information on values is not considered explicitly. One interpretation is that professionals know what their patients want and feed in this information implicitly. Occasionally groups include patients but, given the interests of most of the participants, the discussion tends to revolve around technical matters.

Fink and colleagues (1984) regarded the extraction and synthesis of pertinent information for use by consensus development groups as a major challenge. They suggested that without such reviews participants were more likely to rely on their own particular experiences. Anecdotal evidence suggests that when literature reviews have been provided, the evidence has been used both in deliberation and in the decisions, making the consensus process easier (Jacoby, 1988). The means of providing information to participants varies, from sole reliance on the expertise and knowledge of members of groups to provision of a comprehensive synthesis of relevant research (Lomas, 1991). Sometimes it may be decided not to use formal information because the available evidence is so weak (Merrick *et al,* 1987).

As part of **planning**, the organisers have to decide whether to provide information and if so what to provide and in what form to provide it. Initially we chose to focus on this aspect of 'information' but discovered only two studies which had directly examined the issues. This chapter therefore focuses on **individual judgement** (how individuals deal with information) and **group interaction** (how information is used within groups). Three general questions are addressed.

- How does information influence individual decision-making?
- Does the feedback of individual judgements influence group decision-making?
- How does information influence group decision-making?

How does information influence individual decision-making?

There is a large amount of research, from a variety of areas in psychology (outlined in chapter 2), about how individuals acquire, assimilate, interpret and use information. The research reviewed here concentrates on the area of decision-making and is merely an attempt to highlight some of the issues involved. We focus on the following questions.

- Does the way in which information is presented influence judgements?
- Does the way in which information is framed influence decision-making?
- Do existing beliefs affect the way in which people interpret information?
- How do experts use information in decision-making?

Does the way in which information is presented influence judgements?

The way in which information is presented can influence individuals' judgement (Payne *et al,* 1992). For example, in consumer purchasing decisions Russo (1977) showed that when the unit prices of products (e.g. cost per ounce) were displayed on shelf tags, consumer expenditure decreased. When they were displayed on a list (all grouped together) expenditure decreased further, that is consumers bought more of the less expensive products. Russo argued that this was because the display of information in this way made the information easier to process, thus increasing its use in purchasing decisions.

Information can be displayed according to alternatives or attributes of those alternatives, for example by type of drug, or by attributes of those drugs (side-effects, efficacy, cost). When information is displayed by attribute, processing is more likely to be by attribute and when it is displayed by alternative, processing is more likely to be by alternative (Bettman & Kakkar, 1977; Jarvenpaa, 1990). Thus, the way in which information is displayed can make it more likely that people will use the information and can influence how the information is used.

In a clinical context, the same information can be presented in different ways. For example, there are different methods for summarising the outcomes of clinical trials. Outcomes can be reported as relative risk reduction, absolute risk reduction, proportion of event-free patients, or the number of patients who need to be treated to prevent an event (Laupacis *et al,* 1988).

Research has shown that the way in which the results of clinical trials are presented can influence judgement. Fahey *et al* (1995) asked those responsible for health policy to rate whether they would support purchasing a breast screening programme and a cardiac rehabilitation programme. Data from a single clinical trial on the effectiveness of breast screening and from a systematic review on the effectiveness of cardiac rehabilitation were presented. The results were presented in each of the four ways listed above but the different types of presentation were described as being the results of different trials rather than all being from the same source. The decision to fund the programme was significantly affected by the way in which the information was presented, with reporting of relative risk leading to higher support for purchasing (*Table 14*). Only three out of 140 respondents stated that they realised that all four results summarised the same data.

In a similar study, Bucher and colleagues (1994) investigated the judgements of Swiss hospital doctors and general practitioners on drug treatment in lowering serum cholesterol levels. Respondents completed one of two questionnaires. One questionnaire reported the results of a clinical trial as relative risk reductions and the other reported them as absolute risk reductions for three outcomes: non-fatal and fatal myocardial infarction combined, fatal myocardial infarction alone, and total mortality. Respondents rated the effectiveness of treatment and their likelihood of prescribing drug treatment. The method of summarising the results of trials influenced the respondents' judgements of the effectiveness of treatment and their likelihood of starting treatment. Ratings of effectiveness were higher for relative risk reduction than for absolute risk reduction.

Does the way in which information is framed influence decision-making?

The way in which information is framed can also influence judgement (Tversky & Kahneman, 1981).

TABLE 14 *Mean score on two health policy issues made by health authority members when the same results were presented in four different ways (Fahey* et al, *1995)*

Method of data presentation	Mammography		Cardiac rehabilitation	
	Data presented	Mean score, % (95% CI)	Data presented	Mean score, % (95% CI)
Relative risk reduction	RRR = 34%	79 (76 to 83)	RRR = 20%	76 (72 to 80)
Absolute risk reduction	ARR = 0.06%	38 (35 to 42)	ARR = 3%	56 (53 to 60)
Proportion of event-free patients	EFP = 99.82 vs. 99.8%	38 (34 to 42)	EFP = 84 vs. 87%	53 (49 to 57)
Number needed to treat	NNT = 1592	51 (47 to 55)	NNT = 31	62 (58 to 66)

A higher score indicates stronger support for purchasing the programme
CI = confidence interval

An example of this is the 'Asian disease problem' in which participants are presented with one of two versions of a proposed response to an epidemic that is expected to kill 600 people. In one version, subjects are given a choice between two possible courses of action: programme A in which 200 people are saved or programme B in which there is a one in three chance that 600 people are saved and a two in three chance that no one will be saved. A majority (72%) of participants preferred programme A. In the other version the choices are programme A in which 400 people die, or programme B in which there is a one in three that no one will die and a two in three chance that 600 people will die. In this version the preferences were reversed with most people (72%) preferring B to A. This demonstrates a fairly general finding that people are risk averse for gains and risk seeking for losses. In other words, they prefer a certain good outcome, but are prepared to take a risk when faced with loss. Thus framing a decision problem (or the information provided to make a decision) in terms of gains or losses can alter judgements.

Similar research was conducted by McNeil and colleagues (1982). They presented study participants with a choice between two types of treatment for lung cancer: radiation therapy or surgery and provided expected outcomes of these treatments for three time periods (immediately after treatment, after 1 year and after 5 years). The data given showed that surgery offered higher life expectancy but had a greater risk of immediate death. As in the previous example, the information was framed either in terms of mortality or in terms of survival. Only 18% of subjects favoured radiation therapy in the survival frame, whereas 44% favoured it in the mortality frame. This was the case even for experienced doctors. McNeil and colleagues (1988) extended this study by including a mixed frame (giving both survival and mortality data). Results were similar to the first study for the survival and mortality frames while the mixed frame was in between, but closer to the mortality frame (*Table 15*). This study was conducted among postgraduate students in radiology departments in the USA and was also extended cross-culturally to include Israeli postgraduate student radiologists.

TABLE 15 *Proportion of respondents who favoured radiation therapy rather than surgery for lung cancer under three different formats (McNeil* et al, *1988)*

Format	US respondents (n (%))	Israeli respondents (n (%))	Total (%)
Survival	87 (16)	126 (20)	18
Mortality	80 (50)	132 (45)	47
Mixed	223 (44)	144 (34)	40

Do existing beliefs affect the way in which people interpret information?

A number of studies have shown that people's prior beliefs influence how they interpret new information. Strong beliefs can be very resistant to change even in the face of countervailing evidence (Lord *et al,* 1979). People often ignore such conflicting evidence or reinterpret it to fit with their beliefs (Nisbett & Ross, 1980).

Koehler (1993) has shown how this can occur in the interpretation of scientific evidence. In two studies he found that research reports that agreed with scientists' prior beliefs were judged to be of higher quality than reports which did not. In one experiment involving graduate science students he induced beliefs (either strong or weak) about the correctness of fictitious scientific issues. Subjects then read research reports of either high or low quality, the results of which either supported or opposed the induced prior beliefs. Participants gave higher evaluative ratings to reports that agreed with their prior beliefs, and this was especially so among participants with strong prior beliefs.

In a second study, Koehler sent a hypothetical parapsychological research report to parapsychologists and to scientists affiliated with sceptical organisations. Each person received only one report which was of either high or low quality, the results of which were either in line with or contradictory to the person's prior beliefs. As in the previous study, reports that agreed with the scientists' prior beliefs were evaluated more favourably than those which did not. Thus even in the interpretation of scientific evidence people may be more likely to agree with studies which favour their prior positions.

How do experts use information in decision-making?

The way in which people use information also influences judgement, and experts may not use information appropriately (Payne *et al,* 1992). One example is the use of information from clinical test results to determine the probability of a particular event, such as the presence of disease. Measures such as the sensitivity and specificity of the test, on which such judgements are often made, are them-

selves dependent on the underlying prevalence of the condition (Gigerenzer *et al,* 1988).

Shanteau (1992) reviewed research which compared information use by experts and novices. Although experts might be expected to use more information than novices in making their judgements, there is little empirical support for this. Both experts and novices have been shown to use a similar amount of information – usually less than the amount of information available – in making judgements. Shanteau suggests that what seems to separate expert judgement from novice judgement is the ability of experts to judge what information is relevant and what information is irrelevant in a given context. But expertise is domain-specific; that is, a person who is expert in one area is not necessarily expert in other. So, for example, a doctor who is expert in diagnosing and treating disease, and thus in selecting the relevant information, may not have expertise in judging the quality of clinical trials, and thus may be less able to discern how much weight is appropriate to attach to the results.

Summary

The information presented to individuals is an important element in decision-making. Information can influence judgement in variety of ways. The way in which information is presented can influence the likelihood of its being used and how it is used. The particular way in which information is framed can also influence judgement. People tend to be risk averse for gains and risk seeking for losses. Individuals' own prior beliefs influence their interpretation of new information. Experts may be better than novices at determining what information is relevant but only in those areas in which they have expertise.

Does the feedback of individual judgements influence group decision-making?

As well as providing a review of the literature on the topic to participants, some group methods involve the provision of information about the judgements of other group members. This raises two questions.

- How does feedback influence judgement?
- What type of feedback is best?

How does feedback influence judgement?

Both the Delphi method and the NGT provide feedback to participants. The role of feedback in the Delphi method is paramount, since this is the only communication amongst group members. In the Delphi method, feedback usually includes the distribution of participants' judgements or the mean or median group judgement. Justifications, rationales or other comments may also be included.

The few studies of the Delphi method that have examined the effect of feedback (Woudenberg, 1991) have shown that the dispersion of participants' views lessens with each round of rating (see *Table 16* for an example). It is unclear whether this is the result of the participants having considered their fellow participants' views or is simply the iterative effect of making their judgement again. Given this tendency, the accuracy of the final group decision will inevitably depend upon the accuracy of the initial group judgement.

A study which asked subjects to forecast if and when a number of events would occur attempted to separate the effects of iteration from those of feedback (Parente *et al,* 1984). There were four groups: one was polled only once and received no feedback, one was polled once and received feedback from another group, one was polled twice but received no feedback, and one was polled twice and received feedback. In terms of 'if' forecasts, neither feedback nor iteration led to an improvement in accuracy (defined by whether the predicted events actually occurred). For 'when' forecasts, accuracy increased due to iteration rather than feedback.

There is little evidence about how feedback in the Delphi method affects group performance, though Rowe and colleagues (1991) have offered a theoretical analysis. In most experimental studies of the method there has been only very limited exchange of information, such as the feedback of group mean scores or frequencies. In view of this limited amount of feedback it is likely that any convergence results from a normative rather than informational influence. That is, participants are being swayed by others' positions since no arguments are being put forward and no other information is being introduced. This suggests that although the Delphi method may remove some of the problems of interaction, it also can remove some of the benefits such as the exchange of information.

Does convergence occur in studies of ratings of the appropriateness of interventions using an NGT? To answer this, initial and final group judgements have been compared. Some studies have reported the proportion of scenarios about which the group members agree (*Table 17*). As can be

TABLE 16 *Consultant physicians' judgements of the value of indicators of junior doctor quality (Jones* et al, *1992)*

	Responses to first round			Responses to second round		
Indicator	**Median**	**Mean**	**SD**	**Median**	**Mean**	**SD**
Ability to communicate with other staff	8	7.67	1.35	8	7.77	1.23
Ability to communicate in writing and orally	8	7.60	1.41	8	7.75	1.22
Membership of Royal College of Physicians	7	6.66	1.86	7	6.23	1.92
Country of qualification	6	5.73	2.62	6	6.03	1.92
Possession of MD or PhD	6	5.37	2.57	6	5.48	2.38
Having held appointment in teaching hospital	6	5.30	1.99	6	5.42	1.92
Number of attempts to gain membership	6	5.30	1.99	6	5.42	1.92
Number of publications or case reports	5	4.86	2.25	5	4.91	2.12
Time from qualification to registrar grade	5	4.86	2.25	5	4.91	2.12
Length of time in registrar grade	4	4.19	2.37	4	3.84	2.12

Scoring system: 0 = no support; 9 = total support
SD = standard deviation

TABLE 17 *Change in percentage of agreed-upon indications from initial to final ratings using an NGT*

Reference	**Topic**	**Initial ratings (%)**	**Final ratings (%)**	**Change (%)**
Chassin *et al,* 1986	CABG	23.2	41.9	18.7
Merrick *et al,* 1987	Carotid endarterectomy*	55.6	53.8	–1.8
Coulter *et al,* 1995	Spinal manipulation			
	– mixed panel	11.8	35.7	23.9
	– chiropractic panel	27.2	63.2	36.0
Kahan *et al,* 1996	AAA surgery	55	58	3
	Carotid endarterectomy	48	60	12
	Cataract surgery*	16	52	36
	Coronary angiography	38	40	2
	CABG	29	41	12
	Coronary angioplasty	27	40	13

* *The number of scenarios changed from the first to the second round making comparisons difficult*

seen, there is considerable variation in the amount of change that occurs. Ignoring those studies in which the number of scenarios changed between rounds, some procedures show very little change between the initial and final ratings (2%), some show a modest amount of change (12–13%), while the study by Coulter and colleagues (1995) showed considerable change (36%).

Others have reported the difference between the initial and final mean deviation from the median (*Table 18*). A decrease indicates that the amount of dispersion around the median has lessened and thus there has been convergence. Reported changes vary from –0.14 to –0.64. Thus it appears that convergence may be relatively slight and is rarely very large.

What type of feedback is best?

Few studies have compared different types of feedback. The degree of information included in the feedback might be expected to improve the accuracy of the final group decision. Woudenberg (1991) reported three studies in which a slight increase in the accuracy of a group decision was indeed obtained as a result of feeding back the

TABLE 18 *Change in mean deviation from the median for initial and final ratings using an NGT*

Reference	Topic	Initial ratings	Final ratings	Change
Park *et al,* 1986	Coronary angiography	1.37	0.86	–0.52
	CABG	1.58	1.06	–0.64
	Cholecystectomy	1.29	0.98	–0.31
	Endoscopy	1.39	0.83	–0.56
	Colonoscopy	1.91	1.36	–0.55
	Carotid endarterectomy	0.52	0.34	–0.17
Scott & Black, 1991b	Cholecystectomy			
	– mixed group	1.25	0.95	–0.30
	– surgical group	1.16	1.02	–0.14
Ballard *et al,* 1992	AAA surgery	1.10	1.00	–0.10
Bernstein *et al,* 1992	Coronary angiography	1.53	1.22	–0.31
Matcher *et al,* 1992	Carotid endarterectomy	1.24	0.95	–0.29
Coulter *et al,* 1995	Spinal manipulation			
	– mixed panel	1.70	1.14	–0.54
	– chiropractic panel	1.39	0.83	–0.56

reasons for individual judgements rather than just the ratings. However, in a fourth study the accuracy decreased.

Gowan and McNichols (1993) examined how different forms of feedback affected the extent of consensus in the final round. Using the Delphi method with two rounds, loan officers made judgements about whether firms would be successful 1 year in the future. They were provided with various pieces of information on which to base their judgements. Three different groups were created, each receiving a different form of feedback: descriptive statistics (the frequencies of participants' judgements), policy capturing models (beta weights) and if–then rules (the rules on which judgements were made). Consensus was greatest for if–then feedback and least for descriptive statistics. However, it is unclear whether greater consensus equated with a better judgement.

Summary

Studies often show a convergence of opinion. Thus it appears that feedback does have an effect on judgement. Convergence in ratings of appropriateness varies considerably, from slight to modest. The evidence as to whether convergence through feedback leads to improved judgement is slight. It may be that merely making the judgement again can improve accuracy. There is little research as to what type of feedback is best, though there is some evidence that information on the reasons for divergent views is more useful than simply feeding back the ratings.

How does information influence group decision-making?

In this section we are concerned with the impact of information within consensus development groups on group decisions. Given the important role of information in the consensus process it is surprising how little research has actually addressed this question. Our search of the literature revealed only two studies which explicitly looked at the impact of information on consensus development in the health field.

We firstly examine the limited amount of research from the health sector and then basic research on the influence of information on groups and the use of information in informal groups. The following questions are addressed.

- Does information influence participants?
- What types of influence operate in groups to produce opinion change?
- What type of influence is strongest: informational or normative?
- How is information used in groups?
- Does information which is shared have more influence than information which is not shared?

Does information influence participants?

Vinokur and colleagues (1985) assessed the impact of information on panellists and speakers at several consensus development conferences using questionnaires which asked about the information they to which were exposed (its novelty and comprehensiveness) and how that information affected them (reinforced their views, changed their views). They found a correlation between the reported novelty of the information and the extent of opinion change ($r = 0.77$, $p < 0.001$). There was also a negative correlation between the expertise of the respondent and reported novelty ($r = –0.44$, $p < 0.01$). Those who were more expert reported that the information was less novel. However, there was no significant correlation between the reported impact of the information on the panellists and the quality of the consensus statement (assessed through content analysis of the statements).

The explanation for the lack of a correlation is not very convincing. The authors suggested that speakers, being more expert, were more likely to understand the significance of the information. It is not clear why this should correlate with the quality of the consensus statement for speakers but not for the panellists who produced the statement. This study suggests that the quality of the statement was related to informational impact, the influence of which was not perceived by the panellists. Thus although this study suggests that information does influence panellists, it is not clear how this occurs.

The issue of whether judgements at consensus conferences are based on evidence has been addressed by Lomas and colleagues (1988) in a study of the National Consensus Conference on Aspects of Cesarean Birth in Canada. Panellists received literature reviews covering each of three areas under consideration. After reading these, but before the conference, panellists completed a questionnaire asking them to rate scenarios for the appropriateness of Caesarean section. After the conference the panellists completed the questionnaire again. Lomas and colleagues (1988) assessed the amount of consensus amongst panellists for those scenarios for which there was good research evidence of appropriateness and for those for which research evidence was conflicting, poor or non-existent (*Table 19*).

Three levels of agreement were defined: total agreement (all within a 4-point range at either end of the scale), partial agreement (80% within the 4-point ranges), and disagreement (neither of the above). Before the conference there was greater agreement (partial and total agreement combined) for scenarios with good research evidence than for scenarios that lacked such support (85% and 30%, respectively). Following the conference, levels of agreement had improved in 71% of the research-based scenarios but in only 24% of the other scenarios. Thus Lomas and colleagues (1988) concluded that the consensus process was sensitive to the availability of good research evidence.

However, this may not be the complete picture. Consensus was certainly high for research-based scenarios both before (85%) and after (97%) the conference. In contrast, the level of agreement for scenarios lacking research support rose from only 30% to 38%. This suggests that it is easier to reach agreement when research evidence exists. However, given that scenarios without a research basis were more numerous than research-based ones (192 versus 32), more of the scenarios about which the group agreed had no research basis (73) than had one (31). Clearly the existence of a research basis is not an essential requirement for reaching a consensus.

What type of influence is strongest: informational or normative?

Two main sources of influence occur in groups: informational and normative. Changes in opinion following group discussion can be brought about

TABLE 19 *Level of agreement on research and non-research based scenarios before and after a consensus conference (Lomas* et al, *1988)*

	Research-based		Non-research-based	
	Before (n (%))	**After (n (%))**	**Before (n (%))**	**After (n (%))**
Disagree	5 (16)	1* (3)	134 (70)	119* (62)
Partial agreement	16 (50)	12* (28)	50 (26)	42* (22)
Total agreement	22 (34)	30* (69)	8 (4)	31* (16)

* *These numbers were calculated from percentages given in the paper*

by the exchange of information or relevant arguments within the group. Change may also be brought about by exposure to others' opinions or choices, through the pressure such exposure exerts to conform to normative standards. Much research suggests that informational influence is generally stronger (that is, produces more shift in opinion) than normative influence (Kaplan, 1987). When provided with others' preferences, but not arguments, opinion polarisation is weaker (Clark & Willems, 1969). When provided with information but not preferences, opinion shift is stronger (Burnstein & Vinokur, 1975).

Differences in the type of decision task may affect the type of influence which predominates. Kaplan and Miller (1987) had mock juries make decisions on two types of damage awards: compensatory damages, where documented losses exist and can be factually argued, and exemplary damages, which are intended to punish the defendant and rest on the perceived deviance from standards. The former they regarded as an intellectual, factual-based decision, the latter as a judgemental, normative-based decision. They coded the group's discussion according to the type of influence statements involved and found that for the intellectual decision, informational influence was greater, whereas for the judgemental issue, normative influence was greater. Thus, in general, in tasks for which decisions are perceived to be factually based the influence of information may be greater than normative pressures.

Another question is what type of information is most influential in changing opinion. Vinokur and Burnstein (1978a) suggested that information that is novel is likely to have the most impact in changing opinion. This is in line with their study of consensus development conferences reported earlier (Vinokur *et al,* 1985).

How is information used in groups?

A closer look at how information is used in groups reveals that people tend to offer information which supports their position. Thus if there is a majority position within the group, statements in support of this position are likely to predominate. The content of the discussion does seem to influence individuals' opinions (Kaplan, 1987).

Change in opinion has been shown to be greater for issues with which people are less familiar (Vinokur & Burnstein, 1978b). This leads to the question of how, if group members share the same information (such as a literature review), changes in opinion might be expected (since the initial opinions might have been based on the same information). Kaplan (1987) suggests that in part it is likely that these pre-discussion judgements are based on a subset of the original information, through individuals overlooking, misunderstanding, or not integrating all the relevant information. During discussion, members may share information which was salient to them during their initial judgement and thus the information introduced through interaction will not be identical for all participants.

Does information which is shared have more influence than information which is not shared?

Information may not be shared between group members. For example, when a group is made up of people with different backgrounds, each member may possess different background information. Stasser (1992) suggested that during group discussion individuals sample information from a pool of information which they have available to them, such as past experience. The likelihood of a piece of information entering the group discussion depends, in part, on the number of individuals who possess that piece of information. So, for example, in the context of a discussion of the appropriateness of a healthcare intervention, if all members were aware of a particular piece of research addressing the issue, the likelihood of that piece of research being mentioned would be greater than if only one member was aware of it.

Stasser and colleagues have conducted a number of studies in this area. The basic research design has been to provide information differentially to group members so that some pieces of information are shared by all group members and others are not. Using the problem of selecting a candidate for president of a student organisation they found that, in general, shared information was more likely to be recalled and to be mentioned in discussion (Stasser & Titus, 1987; Stasser *et al,* 1989). Shared information was discussed more when the information load was high and groups were large.

Stasser (1992) also suggested some ways of increasing the use of unshared information. He suggested that in real-world groups in which participants have different specialties, the group members may recognise areas in which each has a unique contribution to make. He also suggested that memory aid techniques may be useful, such as note-taking and the use of flip charts. He recommended the NGT because it incorporates these features as well as allowing all members to contribute.

Schittekatte (1996) replicated and extended some of this work by evaluating ways of increasing the use of unshared information. Making the unshared information more salient and having information shared by one other group member increased its use in discussion. However, more widely shared information still predominated, was more likely to enter discussion, to be reacted to by other group members and to be repeated.

Summary

The small amount of research on the effects of information on consensus groups suggests that information does affect their decisions. Research on basic group processes suggests that both informational and normative influences are likely to operate in groups. Informational influence is often dominant, especially in fact-oriented groups. Confirmatory or supportive information is most likely to be discussed as is information that is shared among group members. Novel information may have the greatest impact on opinion change.

Implications for clinical guideline development

Provision of information

Research-based information should be provided to all participants at an early stage for five reasons.

- It has an impact.
- If all members of the group have access to such information it is more likely to be discussed within the group.
- Providing a literature review to group members before discussion may increase the perception that the task is research-based which will encourage members to be more reliant on information.
- If group members come to the discussion with opinions that are, at least to some extent, based on a reading of relevant research, the information exchanged may be more likely to reflect the research evidence.
- Providing a common starting point may help group cohesion.

Group members should be encouraged to bring the review, and any notes they made on it, to the group sessions as a memory aid.

Presentation of information

Information presented in the form of articles or abstracts is likely to be less easily assimilated than information presented in a synthesised form, such as tables. Information that is presented in a format that is easy to read and understand may be more likely to be used by participants. If the information is tabulated in a way which makes salient the dimensions on which to base judgements it is more likely to be processed in this manner.

Preparation of information

Although there is no scientific research to demonstrate the value of involving methodologists, it seems sensible that they should be involved in conducting any literature review, since they are expert in judging the quality of research. Clinicians, on the other hand, may not be expert in these judgements. Organisation by methodologists may give precedence to those factors which are relevant for making a judgement over those which are irrelevant, thus making it more likely that judgements will be based on the appropriate information.

Grading the quality of studies may mitigate the biases of the reviewers somewhat, but may not eliminate them.

Feedback of individual judgements

With NGTs and Delphi methods, two or more rating rounds are likely to result in some convergence of individual judgements, though it is unclear whether this increases the accuracy of the group decision.

Although there is little research on the value and impact of feedback, especially with the Delphi method, it may be advisable to feed back reasons or arguments as well as measures of central tendency or dispersion.

Chapter 7

Methods of structuring the interaction

Introduction

During the **planning** phase, the particular consensus development method to be used needs to be selected. In addition, a detailed description of how the method is to be implemented is needed, given the wide variety of applications reported in the literature. If the chosen method does differ from the original application, some justification for this deviation is needed.

Before being asked for their **individual judgements**, the participants will receive information about the process which has been adopted. They are likely to form views on this. Do they think the method is appropriate? Do they understand what they are meant to do? Some people may have previous experience of the method, others may have read about it or have other information about it which may influence their opinion. Those who have used the method before may react differently to those who have not. For example, Stephenson and colleagues (1982) found that those who had used an NGT before were less satisfied with it than those who had no previous experience.

For methods that involve face-to-face interaction (NGTs and consensus development conferences), the way a meeting is structured and organised will affect the way in which a group interacts (Pavitt, 1993; Roth, 1994), and can also influence how a group arrives at a judgement and how the output is produced. In addition, the setting for any meetings may influence the group's judgements, as may the way the meeting is run. It is therefore necessary that the impact of such factors is considered.

How the chosen method affects the production of the output will be considered in chapter 8. Here we focus on issues concerned with planning the consensus development method and on group interaction and address three questions.

- Does the choice of consensus development method influence the group's decision?
- Does the setting for the group meetings affect the consensus decision?
- Do the characteristics of a group facilitator affect the consensus decision?

Does the choice of consensus development method influence the group's decision?

The Delphi method and the NGT are the methods most commonly used in the development of guidelines (and are indeed the most commonly used in other areas) and thus the focus will be on these. Sixteen studies were identified which compared the NGT, Delphi method and informal methods, using measures of decision quality as the outcome. A further three studies compared the Delphi method or the NGT with another formal method (*Table 20*). No comparative studies involving consensus development conferences were found.

The studies report a variety of different comparisons with no two studies the same. In addition to differences in the type of task and the methods compared, the studies also differ in the particular way they operationalise the methods used. The operationalisation of the method can vary so widely that some studies which report using an NGT deviate so much from the standard format that we do not consider it to be an NGT (for example, Hegedus & Rasmussen, 1986). We have, therefore, used our own labels to define what type of consensus method was used. This labelling is based on what we consider to be the primary features of each method and acceptable variations, as described in *Table 21*.

Clearly, when comparing methods some means of judging success is needed. As discussed in chapter 3, determining the validity of consensus development methods is problematic. The appropriate means of validating a method depends largely on the type of task being undertaken. In this review, the most relevant measure of validity for the specific task being carried out has been used. Details of the studies can be found in appendix 2.

NGT versus informal methods

Of the ten studies that have compared the NGT with informal groups with regard to decision quality, five found the NGT better than informal groups (Brightman *et al*, 1983; Gustafson *et al*, 1973; Herbert & Yost, 1979; Van de Ven & Delbecq, 1974; White *et al*, 1980), four found

TABLE 20 *Studies comparing methods, categorised by task*

Reference	Comparisons*	Results	Task
Brightman *et al,* 1983	Informal = t NGT = e-f-t-e	NGT best	Probability estimation
Fischer, 1981	Informal = t-e NGT = e-t-e Delphi = e-f-e Staticised = e	No difference	Probability estimation
Gustafson *et al,* 1973	Informal = e NGT = e-t-e Delphi = e-f-e Staticised = e	NGT best	Probability estimation
Larreche & Moinpour, 1983	Informal = e-t Delphi = e-f-e Staticised = e	Delphi best	Forecasting
Sniezek, 1990	Informal = e-t Delphi = e-f-e Staticised = e	No difference	Forecasting
Soon & O'Connor, 1991	Informal = t NGT = e-t-e Staticised = e	No difference	Forecasting
Boje & Murnighan, 1982	NGT = e-f-t-e Delphi = e-f-e Staticised = e	No difference	Probability estimation and almanac questions
Dalkey, 1969	Informal = t Delphi = e-f-e	Delphi best	Almanac questions
Felsenthal & Fuchs, 1976	Informal = t-e NGT = e-t-e Delphi = e-f-e Staticised = e	3-person groups: no difference 6-person groups: Delphi best	Almanac questions
Burleson *et al,* 1984	Informal = e-t Delphi = e-f-e Staticised = e-e	Informal best	Ranking
Erffmeyer & Lane, 1984	Informal = e-t NGT = e-f-t-e Delphi = e-f-e Structured = e-t	Delphi best NGT worst	Ranking
Herbert & Yost, 1979	Informal = e-t NGT = e-f-t-e	NGT best	Ranking
Nemiroff *et al,* 1976	Informal = e-t NGT = e-f-t-e Structured = e-t	No difference	Ranking
Jarboe, 1988	NGT = e-f-t-e Structured = t	NGT best	Idea generation
Van de Ven & Delbecq, 1974	Informal = t NGT = e-f-t-e Delphi = e-f-e	NGT and Delphi better than informal	Idea generation
White *et al,* 1980	Informal = t NGT = e-f-t-e Structured = t	NGT best	Idea generation
Miner, 1979	NGT = e-f-t-e? Delphi = e-f-e?	No difference	Role-play task
Rohrbaugh, 1979	NGT = e-f-t-e SJA = e-f-t	No difference	Judgement policy
Rohrbaugh, 1981	Delphi = e-f-e SJA = e-f-t	No difference	Judgement policy

* *See* Table 21

TABLE 21 *Essential features of consensus development methods and variations in how the methods are operationalised*

Method	Operationalisation
Informal methods	talk (t) estimate – talk (e-t) talk – estimate (t-e)
NGT	estimate – feedback – talk – estimate (e-f-t-e)* estimate – talk – estimate (e-t-e)
Delphi method	estimate – feedback – estimate (e-f-e)
Staticised groups	estimate (e) estimate – estimate (e-e)

* *The standard NGT; the round-robin format is considered as feedback, but other types of feedback may also occur.*

Note that an 'estimate' is shorthand for the individual judgement and will vary depending on the type of task. For example, it may be ranking a number of items or generating ideas. The essential point is that the individual makes the estimate.

no difference (Felsenthal & Fuchs, 1976; Fischer, 1981; Nemiroff *et al,* 1976; Soon & O'Connor, 1991) and one found the NGT worse than interacting groups (Erffmeyer & Lane, 1984).

Part of the problem in generalising from these studies lies in the different ways in which the methods, including the informal methods, were operationalised. There are differences between the studies in the method of feedback, the structure of the interaction, the use of facilitators, and the number of rounds used. Of the five studies which found the NGT to be better, four used a round-robin format to provide feedback (Brightman *et al,* 1983; Herbert & Yost, 1979; Jarboe, 1988; White *et al,* 1980) and one did not (Gustafson *et al,* 1973). Three of the four studies that used a round-robin format had trained group leaders (Brightman *et al,* 1983; Jarboe, 1988; White *et al,* 1980), whilst one report was unclear on this (Herbert & Yost, 1979). All four included an opportunity for discussion, but differed in the conduct of the NGT. Of the four studies which found no difference between NGT and informal methods, two did not maintain the round-robin format (Fischer, 1981; Soon & O'Connor, 1991) and the other two had groups which had no leader and had to organise themselves (Boje & Murnigham, 1982; Nemiroff *et al,* 1976). The one study that found the NGT to be worse adhered to the standard NGT procedure (Erffmeyer & Lane, 1984).

Thus, it is likely that the particular way the procedure is carried out is important. It is often difficult to determine from the literature exactly how the procedure was implemented, but in general in studies which used facilitators and which stayed closest to the original format, the NGT tended to perform better.

Is the success of a particular consensus development method partly dependent on the type of task being undertaken? For most types of tasks, some studies have shown the NGT to be better and some have not, and so it is not clear that the specifics of the task interact with the method, though some studies suggest that the level of task difficulty is important. Some commentators have suggested that an NGT is better for idea generation (Hegedus & Rasmussen, 1986), though some of the studies they considered are ones which diverge greatly from the NGT format.

Delphi versus informal methods

Of the studies that compared the Delphi method with informal methods, nine used measures of accuracy. Four found the Delphi method was better (Dalkey, 1969; Erffmeyer & Lane, 1984; Larreche & Moinpour, 1983; Van de Ven & Delbecq, 1974), three found the Delphi method to be equivalent to informal methods (Fischer, 1981; Gustafson *et al,* 1973; Sneizek, 1990), one found that a variation on Delphi was worse (Burleson *et al,* 1984), and one found the Delphi method was better for groups of six but no different for groups of three (Felsenthal & Fuchs, 1976).

The performance of the Delphi method may vary by task to some extent. Of the four studies which examined probability estimations and forecasting tasks, three found the Delphi method no better than informal groups (Fischer, 1981; Sneizek, 1990; Soon & O'Connor, 1991), while only one found it better (Larreche & Moinpour, 1983). Of the three which used almanac questions, two found the Delphi method better (Dalkey, 1969; Felsenthal & Fuchs, 1976) and one did not (Boje & Murnigham, 1982). Of the two which used ranking tasks, one found the Delphi method better (Erffmeyer & Lane, 1984) and one found it worse than informal methods (Burleson *et al,* 1984). The one study which used idea generation found the Delphi method better than informal methods (Van de Ven & Delbecq, 1974).

The success of the Delphi method also depends on the way in which the method is implemented. For example, success may depend on the following factors.

- **The number of feedback rounds.** Most studies had only one round of feedback. Of the four studies that had three to six rounds, two found the Delphi method better than informal methods (Felsenthal & Fuchs, 1976; Larreche & Moinpour, 1983) and two did not (Boje & Murnigham, 1982; Miner, 1979).
- **The type of feedback given.** Only three studies gave more feedback than just the estimate or solution (Burleson *et al,* 1984; Erffmeyer & Lane, 1984; Larreche & Moinpour, 1983). Two of these studies found the Delphi method better than informal groups (Erffmeyer & Lane, 1984; Larreche & Moinpour, 1983) and one found it worse (Burleson *et al,* 1984). Rowe (1992) has suggested that when only group estimates such as means or medians are fed back, there is little room for change due to informational influences but more likelihood of change due to normative influences. This may be less so when the task is idea generation, where the ideas are fed back.
- **The physical location of participants.** Van de Ven & Delbecq (1974) and Erffmeyer & Lane (1984) were the only researchers to conduct each round

exclusively by mail. Both found the Delphi method was better than informal groups. Five studies had participants in the same room, with no direct communication between them but with varying degrees of anonymity. Being in the same room, and communicating by notes, or communicating with a person in the next room or behind a partition, may have an adverse effect on the development of consensus.

NGT versus Delphi method

Seven studies compared NGTs and the Delphi method. One found that the NGT was better (Gustafson *et al,* 1973), two found that the Delphi method was better (Erffmeyer & Lane, 1984; Felsenthal & Fuchs, 1976) – though for one of these studies (Felsenthal & Fuchs, 1976) only for six-person and not three-person groups (Felsenthal & Fuchs, 1976) – and four found no differences (Boje & Murnighan, 1982; Fischer, 1981; Miner, 1979; Van de Ven & Delbecq, 1974). Again, there is no clear pattern as to what type of tasks or particular aspects of the procedure might be more or less important in producing these differences.

Other formal consensus development methods

As has been seen in chapter 1, there are various other formal methods for achieving consensus. Most have been investigated rarely. Some of these methods have been compared with the NGT and with the Delphi method.

Structured discussion methods structure the interaction so that discussion proceeds in a logical manner. White and colleagues (1980), who used trained facilitators to structure the way questions were discussed, found this method inferior to the NGT for simple problems, but no different for moderately difficult or complex problems. Structured interaction was better than informal methods for moderately difficult or complex problems, but no different for simple problems. Jarboe (1988) also found structured discussion inferior to the NGT. Other studies have also shown that giving instructions on how to reach consensus leads to better performance than informal groups (Hall & Watson, 1970; Innami, 1994; Nemiroff & King, 1975).

In separate studies, Rohrbaugh (1979; 1981) compared SJA with the Delphi method and the NGT. SJA did not lead to better quality group judgement than either the Delphi method or the NGT, but it did improve the quality of individual judgement, especially when compared with Delphi groups.

Some commentators have suggested that an overemphasis on consensus may actually decrease the quality of group judgement. They have gone so far as to suggest the use of techniques designed to create constructive conflict (such as the Devil's advocate and dialectical inquiry techniques). Studies comparing these techniques have had mixed results (Priem *et al,* 1995; Priem & Price, 1991; Schweiger *et al,* 1986; Schweiger *et al,* 1989; Schwenk & Cosier, 1993).

Summary

Formal methods generally perform as well or better than informal methods but it is difficult to tell which of the formal methods is best. Formal techniques are said to work because they provide structure to the interaction, though which aspect of the structure is the most important is less well understood. Most studies have not examined whether the interaction is actually altered in the ways suggested, and many studies did not operationalise the technique in a consistent way. Hence, it is difficult to decide which formal technique performs best. It may well be that the particular operationalisation of the technique and, probably, the particular aspects of the task to which it is applied affect the relative success of different formal techniques.

Does the setting for group meetings affect the consensus decision?

Every meeting takes place in some setting, generally a room in some institution. There are often preliminaries before the meeting, such as introductions and the provision of refreshments, to make the participants feel comfortable. The surroundings and general environment for the meeting may be an important aspect in the success of a decision-making meeting.

Reagan-Cirincione and Rohrbaugh (1992), in describing arrangements for a decision conference, stressed the importance of fitting the design and furnishings of a room to the group. They claimed that this is essential for enhancing interpersonal communication and creative thinking, and suggested a number of advantageous features for the layout of the room: square spaces are better than rectangular, it is important to use whiteboards, and chairs should be comfortable and set in a semi-circle (with no tables) to allow complete interaction and a focus on the board.

Findings from other studies support the suggestion that seating arrangements may be important. For

example, people are more likely to talk to those with whom they have eye contact (Argyle & Dean, 1965), and circular seating arrangements lead to more participation than alternative arrangements such as rows (Sommer & Olsen, 1980).

There is little empirical work which has examined the effect of the environment on the quality of group decision-making. Some studies have looked at the effect of stressful environments based on the hypothesis that as the level of stress increases group members are more likely to defer to authority. However, much of this work has been carried out in environments that are unlike those likely to be encountered in the development of clinical guidelines. Driskell and Salas (1991), for example, studied navel cadets in whom stress was induced by suggesting that tear gas might be introduced during the task! Few studies have looked at more mundane, typical types of stress.

Worchel and Shackelford (1991) compared decision-making, with and without leaders, in a conducive environment with that in a noisy or crowded setting. They found that participants rated various aspects of their group and the decision-making more highly when working in a positive rather than a negative environment. However, there was no indication of whether the environment affected their judgements.

Time pressure may affect decision-making. Karau and Kelly (1992) examined the effect of time pressure on the way information is considered by the group. Their results suggest that when time pressure is high, groups focus on completing the task, which can lead to their initial preferences having more influence on both the group discussion and decision. With moderate time pressure, groups focus more on the quality of the output, and attend more carefully to the information.

The mood among participants may also affect decision-making. Isen and Means (1983) showed that those in a positive engaged in less thorough information use than those in a neutral mood. Positive affect, however, may be advantageous for creative problem-solving.

Summary

The environment in which the decision-making session takes place may affect the interaction and satisfaction of participants, and may ultimately have an impact on decision quality. There is little research which actually looks at this question. However, of the many factors which can influence decision-making, except for extreme environments, the environment is likely to have only a marginal impact.

Do the characteristics of a group facilitator affect the consensus decision?

The use of facilitators and the type of facilitators used may be important determinants of group decisions. Vinokur and colleagues (1985) and Wortman and colleagues (1988) studied consensus development conferences, using non-participant observation, self-administered questionnaires for panellists and speakers, and content analysis of the consensus statements to provide data on the quality of the decision-making. Their findings suggested that a facilitative chairperson is one of the most important ingredients in a successful conference. They suggested that this is because the chairperson, through regulation of the interaction and decision procedure, facilitates the exchange of relevant information. This may explain why studies have found that the NGT, which involves a trained facilitator, is superior to informal groups despite the latter requiring effective chairing.

As was mentioned in chapter 2, there is a great deal of research on the effectiveness of leadership. Various approaches, focusing on different aspects of leadership, have been used. Some have concentrated on the behaviour of leaders, often drawing distinctions between socio-emotional types of behaviour and task-oriented types of behaviour. Korman's (1966) review of this work suggested these behaviours are not consistently related to group performance. Fiedler (1967) in his contingency theory of leadership suggested that effective leadership involves an interaction between the leader and the situation, with different types of leadership being effective in different situations. There is some empirical support for this theory (Strube & Garcia, 1981).

Although research in this area covers a range of types of leadership, the majority of it focuses on a type of leader who must lead (the 'boss'), rather than the type of leader who is simply there to facilitate the smooth running of a meeting. Janis's (1982) work on groupthink dealt with the former type of leader, and much work has focused on military leaders, sports team captains, and others with responsibility beyond simply conducting a good meeting. Thus much of the work on leadership is not directly relevant to facilitating or chairing a meeting. However, some findings are relevant

and suggest that leaders can affect the outcome of group decision-making.

Flowers (1977), for example, showed that groups generate more alternatives when leaders encourage members to present diverse opinions than when they encourage consensus. Others have shown that if the leader gives an opinion, then the timing of that opinion can affect the group process. Groups with leaders who delay giving their opinions, rather than state their opinions at the start of the discussion, have been shown to be more productive in generating options (Anderson & Balzer, 1991; Flowers, 1977; Maier & McRay, 1972; Maier & Sashkin, 1971).

The role of the facilitator and what a good facilitator does has received less attention. More recently, with the advent of computer-assisted group decision support systems, the role of the facilitator has begun to receive some attention, though often in these cases the facilitator's role is expanded to include facilitating the use of the support software. Clawson and colleagues (1993) attempted to map the dimensions involved in the role of the facilitator. The facilitator has a range of roles including providing structure for the group interaction, maintaining the agenda, recognising speakers, focusing the group on the outcome, managing conflict, and creating a positive environment. However, there is little evidence as to whether groups can manage equally well without facilitators, and whether there are better or worse ways of facilitating a meeting. In one study George and colleagues (1992) found that the presence of a facilitator led to higher quality decisions, but Anson and colleagues (1995) found that, while a facilitator may not enhance the decision performance of a group, he or she may produce a positive effect on the processes and cohesion of a group.

Summary

Although work on leadership suggests that aspects of the leader's behaviour are important for group decision-making, the models of leadership used are often not directly transferable to facilitation. There is little work which examines what a good facilitator is and very little work which looks at the effects of facilitation on group decision-making. However, it is likely that this key role will influence group decision-making.

Implications for clinical guideline development

Formal versus informal methods

Formal methods (the NGT and the Delphi method) generally perform better than informal ones and thus may be better for consensus development. Although the reasons why they perform better are not clear, it is likely that staying closer to the original format provides better results. Some aspects that are likely to be important include (1) ensuring that all members have a chance to voice their views, (2) ensuring that all options are discussed, (3) providing feedback and repeating the judgement, and (4) ensuring that judgements are made confidentially.

Physical environment for NGTs

A comfortable environment for meetings is likely to be preferred by participants and to be conducive to discussion. There is, however, a lack of scientific evidence to demonstrate this.

Facilitators in NGTs

It is likely that a good facilitator will enhance consensus development but there is no rigorous evidence to support this. A good facilitator can ensure that the procedure is conducted properly.

Chapter 8

Outputs: methods of synthesising individual judgements

Introduction

The nature of the outputs of a consensus development process will be determined during its **planning**, along with decisions as to whom the outputs are aimed. The process of arriving at these decisions will reflect and often help to refine the objectives of the exercise. There may be further revisions when **individual judgements** are being made in the light of participants' opinions about what they are being asked to do. The focus of this chapter is at the stage of **group interaction**, and is on the processes by which individual judgements can be combined or summarised to form the group's output. Various options are available, each of which can be characterised on three dimensions:

- aggregation rules determined and put into practice by a group versus aggregation rules externally imposed and executed
- implicit methods versus explicit methods of aggregation of individuals' judgements
- dichotomous judgement (agree or disagree) versus scaled judgement demonstrating central tendency and extent of agreement within the group.

The choice of method will depend on whether the task is restricted to identifying a set of judgements that enough of those involved are willing to accept (such as the deliberations of a jury), so that the only interest is in whether there is consensus, or whether the aim is also to identify the **extent** to which consensus exists (as generally occurs in NGTs). The former task begs subsidiary questions about the definitions of 'enough' and 'accept'; the latter begs additional questions about how to characterise the extent of disagreement.

In informal consensus development groups, such as committees, the process can be characterised as personal, publicly declared judgements of a dichotomous nature which are aggregated using implicit methods involving simple rules which have been chosen and executed by the group members. The amount of structure that is imposed on the process of arriving at a conclusion is relatively small. Generally, a chairperson summarises participants' views and proposes a group judgement. There may be more or less pressure on dissidents to come into line. Although many guidelines have been produced by informal groups, one of the problems with these is that it is not clear to the outsider what processes were involved, to what extent divisions of opinion were played down, and how much pressure the participants were under to conform. Justice may have been done, but it may not be seen to have been done.

In contrast, in formal consensus development methods, individual judgements may remain confidential, aggregation is by explicit methods including some indication of the individuals' degree of agreement, and the group's deliberations are controlled by externally imposed rules. The procedures involved in aggregation fall into two distinct stages. In the first, differential weights may be attached to the contributions of the various participants. (Weights may be fixed for all questions, or vary from one question to another.) In the second, group scores and indices of spread are calculated from each set of participants' assessments, weighted or otherwise.

Five questions need to be addressed.

- Is an implicit approach to aggregation of individuals' judgements sufficient?
- Should the judgements of participants be weighted when using an explicit method?
- How should individuals' judgements be aggregated for any one scenario in an explicit method?
- How should group agreement be defined?
- How should group agreement be measured over many scenarios?

Is an implicit approach to aggregation of individuals' judgements sufficient?

Much of the research on implicit aggregation has been primarily concerned with ranking options, rather than the point and interval estimation which has usually been the task for groups determining

clinical appropriateness. However, there is no reason in principle why such groups should not be asked to express preferences between treatment alternatives, and some of the research results will be discussed.

The most familiar application of this approach is in voting, for which various methods for converting sets of individual preference orderings into group orderings have been proposed and extensively examined. There have been three lines of research. One is essentially **descriptive**, the objective being to discover hidden or unconscious rules in apparently implicit processes. This involves treating the group decision-making process as a kind of 'black box' in which the inputs are the individuals' pre-group judgements and the outputs are actual group decisions. The task is then to infer from the inputs and outputs what the decision rules could have been. The work by Davis and colleagues (1973) on social decision schemes provides examples of this approach. In studies of juries they have shown that when an initial majority exists, the majority decision is likely to be the final outcome. They also found that different types of task invoked different types of hidden decision rules.

The second approach is **experimental**. It involves imposing different decision rules on a group, such as majority voting or a consensus scheme, and examining the effect on decision outcomes. In research on juries it has been found that decision rules which require unanimity lead to more hung juries than simple majority rules (Davis, 1980). Kameda and Sugimori (1993) found that groups that were guided by a unanimity rule were more likely to remain committed to their initial decisions than groups using a majority decision rule.

Some approaches to consensus require a vote to be taken, but the very act of voting can affect the outcome. Davis and colleagues (1993) examined the effect of voting on jury decision-making and found that voting, as opposed to deliberation only, increased the likelihood of hung juries but also led to larger awards of damages. They also found that the timing of the vote can affect outcome. When voting occurred early in the deliberations juries were less likely to be deadlocked than juries that voted later, although those who voted later generally awarded higher damages. Davis and colleagues (1989) also showed that in a open ballot the order in which people voted could affect the outcome.

The third approach is **theoretical**: what is the **best** way of mapping individual preferences into a group preference? 'Best' in this context has tended to reflect the interest of the political scientist in the fairness of a process in representing the views of the participants, not the accuracy of the outcome. This may be relevant, however, to the formation of guidelines when they involve prioritisation.

A variety of paradoxes have been identified in voting systems.

- A set of transitive individual preferences ('A better than B' and 'B better than C' necessarily implies 'A better than C') can lead to an intransitive group preference (a cyclic majority).
- Gaming tends to result in the formation of alliances among the participants, along the lines of 'I will vote for your favourite if you will vote for mine'.
- Arrow (1963) proposed a set of apparently reasonable requirements to ensure fairness and feasibility, and then proved that no collective choice rule could be found that satisfied them all simultaneously. This influential work has since been extended (e.g. Coleman, 1982).

Summary

Voting is only suitable for choosing, ranking or prioritising options rather than assigning specific values to each scenario or statement. Research in this field suggests that the more demanding the rules, the less likely a consensus will be achieved. The very act of voting and the timing of a vote can affect the group decision. Voting systems are subject to several paradoxes that might undermine the validity of the outcome.

Should the judgements of participants be weighted when using an explicit method?

If the views of some participants are more important or more accurate than others, so the logic runs, then perhaps judgements should be differentially weighted to reflect these differences. Research on this has been mainly in the fields of forecasting (Granger, 1989) and estimation of subjective probability distributions (Genest & Zidek, 1986). An aim is to determine which system of weights produces the most accurate results.

The problem though, in real groups, is to determine the appropriate weights as this involves estimating how accurate different participants are likely to be. Ferrell (1985) suggested that weighting for expertise may be helpful for large groups with a

history of interaction, in which different members have different areas of expertise and the task requires a wide range of knowledge. Where members have similar levels of expertise, weighting can be expected to have less effect.

The difficulty is that inappropriate weightings can result in less accurate outcomes than no weightings at all. Flores and White (1989) looked at stock market forecasts by business students using four methods: equal weighting, weighting based on past accuracy, weighting based on self-rated expertise, and weighting based on self-rated confidence in judgements. The lowest mean absolute percentage error was found for weighting based on past accuracy, whereas equal weighting and weighting based on self-rated confidence produced the most error. Weightings based on self-rated expertise were in between. However, the differences between methods was small.

Rowe (1992) suggested that measures of expertise based on past performance may provide a better basis for weighting than self-reporting. However, most studies have found that there is little to be gained from weighting.

Summary

Although weighting by expertise may seem attractive in theory, the benefits are uneven and hard to predict. In practice, it is unclear how weightings should be assigned to different participants. Inappropriate weightings may be worse than no weightings.

How should individuals' judgements be aggregated for any one scenario in an explicit method?

When the task is to summarise or filter a set of individuals' values or scores, weighted or otherwise, the obvious first step is to place the scores in a frequency distribution. This involves minimal loss of information and avoids arbitrary judgement, provided that the scores are unweighted.

Unless all the scores happen to be identical, any arithmetical manipulation that goes beyond this first step will involve a choice of method and thus an element of arbitrary judgement. How far beyond this any aggregation method goes will depend on whether the objective is to identify statements about which there is, rather than is not, a consensus, or whether the interest is in the nature and extent of consensus.

- If the task is to identify those scenarios for which there is a consensus, this begs the question of how to define when consensus exists (see next section). The more demanding the criteria, the more anodyne the statement will be, so that if the requirement is too demanding, either no statements will qualify or those that do will be of very little interest.
- When the task is to identify the nature and extent of consensus, this is more akin to the orthodox problem of statistical summary or estimation of central values and the amount of spread around them.

In both cases, given that results may depend on methods, how should a method be chosen? One criterion is **robustness**, in the sense of lack of sensitivity to outliers or rogue participants. Another is **accuracy**: do some methods of aggregation produce more accurate outcomes than others? In principle, **fairness** could also be a criterion, but it has received relatively little attention in this context.

Early studies using the Delphi method used median and interquartile ranges to characterise the frequency distributions of participants' scores. As long as there are eight or more participants and the distribution is not markedly bimodal, these statistics have the advantage of robustness in the sense of being independent of each extreme value and less sensitive to skew in the distribution of responses.

Huber and Delbecq (1972) used Delphi group judgements to compare aggregation rules with groups of different sizes and with scales with different intervals (*Table 22*). For estimates of the value of known quantities, they found that using the mean of group members' ratings resulted in less error than using the midpoint of the interval chosen by the majority. Using tasks with correct answers, others have also shown calculating the mean to be more accurate than other types of aggregation techniques in a variety of areas (Hogarth, 1978).

One interpretation is that individuals' judgements are subject to random error, which can be averaged out, in which case group sizes of eight to 12 are sufficient (Hogarth, 1978). However, this relies on the assumption that individual judgements are unbiased, which is untenable given the considerable literature demonstrating bias in individual judgement (Tversky & Kahneman, 1974). If there is non-random error (bias) in judgements this will not be eliminated through aggregation, though the elimination of random error will still produce an estimate that, in many cases, is more accurate than that of any individual (Rowe, 1992).

TABLE 22 *Expected absolute error using the arithmetic mean of responses from a continuous scale and using the midpoint of the interval chosen by the majority from a scale of ten intervals (Huber & Delbecq, 1972)*

Judgemental accuracy		Number of judges				
		1	3	5	10	20
SD 5%	Mean	3.9	2.3	1.8	1.3	0.9
	Majority	4.4	3.5	3.1	2.8	2.7
SD 10%	Mean	7.5	4.4	3.4	2.5	1.9
	Majority	7.6	5.7	4.9	3.9	3.0

Summary

The appropriate method of aggregating individual judgements will depend on whether the objective is to identify consensus (rather than a lack of consensus) or the nature and extent of consensus. Use of a frequency distribution avoids arbitrary value judgements. Criteria for choosing a method include robustness, accuracy and fairness. In other contexts, the median and inter-quartile range have been shown to be robust.

How should group agreement be defined?

The previous section addressed the issue of how to summarise a group's response for any one scenario. The next question is how to measure the amount of agreement within a group over the consensus exercise for a whole set of scenarios.

One approach involves taking the summarisation process for each scenario to extremes by transforming the distribution of responses to a dichotomous scale (group agrees/disagrees). Plainly there are many ways in which this can be done, and this section describes the effects of different 'definitions' or transforming algorithms on overall levels of agreement found. This section is confined to analyses in which the simplest indicator of overall agreement, the proportion of scenarios for which the group agrees, was used.

In the healthcare literature, two studies have reported the effects on levels of agreement of varying the definitions of consensus, and there is another, unpublished, study. All three studies used the RAND approach (a modified NGT) in which participants' views were expressed on a 9-point scale. Park and colleagues (1989) examined agreement within nine-member groups for three procedures. Scott and Black (1991b) studied two groups – a six-member group of mixed specialties and an eight-member group of surgeons – who rated indications for cholecystectomy. K Imamura and colleagues (personal communication, 1995) studied two groups of orthopaedic surgeons, one in the UK and one in Japan, who rated indications for total hip replacement.

The studies compared a strict definition of agreement (all ratings in the range 1–3 = agreed inappropriate; all ratings in the range 4–6 = agreed equivocal; or all ratings in the range 7–9 = agreed appropriate) with a 'relaxed' definition (all ratings within any 3-point range). They also examined the effect of excluding two extreme ratings. Park and colleagues excluded the maximum and minimum ratings. While the other two studies did that, they also examined the effect of excluding the ratings most distant from the median regardless of direction. In all three studies, relaxing the definition of agreement had little effect on the amount of agreement (*Table 23*). Excluding outliers, however, had a substantial impact.

The studies went on to examine levels of disagreement. 'Strict' disagreement involved at least one rating of 1 and at least one of 9; 'relaxed' disagreement involved one rating in the 1–3 range and one in the 7–9 range (*Table 24*). Not surprisingly, relaxing the definition in this way resulted in increased disagreement. Again, eliminating outliers had a marked effect.

Naylor and colleagues (1990) obtained similar findings when investigating the effect of excluding outliers. They examined different definitions of consensus within a group of 16 people who rated 438 scenarios for acceptable delay before coronary re-vascularisation. Ratings of the urgency of treatment were made on a 9-point scale, with 1 indicating an emergency and 7 indicating marked delay; points 8 and 9 were non-intervention ratings. If agreement was defined as all participants agreeing on a single point, then there was no agreement. If agreement of 75% of participants was regarded as sufficient, then there was agreement on 1.4% of the scenarios. If a simple majority was sufficient, there was agreement on 23.2% of the scenarios.

TABLE 23 *Impact of different definitions of agreement on the level of agreement (%)*

		Include all ratings		Exclude furthest ratings		Exclude min–max ratings	
Reference	**Topic**	**Strict**	**Relaxed**	**Strict**	**Relaxed**	**Strict**	**Relaxed**
Park *et al,* 1989	Coronary angiography	28	29	–	–	50	56
	Endoscopy	25	25	–	–	41	42
	Carotid endarterectomy	41	41	–	–	53	54
Scott & Black, 1991b	Cholecystectomy						
	– mixed panel	45	47	63	67	–	–
	– surgical panel	35	35	57	61	50	53
Imamura *et al,* 1997	Total hip replacement						
	– Britain	42	42	53	59	48	52
	– Japan	23	32	50	69	44	55

TABLE 24 *Impact of different definitions of disagreement on the level of agreement (%)*

		Include all ratings		Exclude furthest ratings		Exclude min–max ratings	
Reference	**Topic**	**Strict**	**Relaxed**	**Strict**	**Relaxed**	**Strict**	**Relaxed**
Park *et al,* 1989	Coronary angiography	2	30	–	–	0	11
	Endoscopy	30	49	–	–	7	29
	Carotid endarterectomy	15	34	–	–	2	18
Scott & Black, 1991b	Cholecystectomy						
	– mixed panel	10	31	3	15	–	–
	– surgical panel	2	26	0	8	0	11
Imamura *et al,* 1997	Total hip replacement						
	– Britain	0	17	0	1	0	3
	– Japan	0	25	0	5	0	10

Naylor and colleagues (1990) also examined the effect of relaxing the definition of consensus so that agreement was defined as all participants being within a range which represented the amount of delay in treatment and whether treatment was needed (1–4, treat; 5–7, wait; 8–9, do not treat). The proportion of scenarios about which the entire group agreed was 11%, whereas about 75% of the group agreed about 59% of scenarios.

Summary

The simplest indicator of overall agreement is the proportion of scenarios for which a group agrees. The definition of agreement or the transforming algorithm used will affect the amount of agreement obtained. Relaxing the definition of agreement has generally been found to have little effect on the amount of agreement, whereas the exclusion of outliers has a substantial impact. In contrast, relaxing the definition of disagreement has a marked effect on the amount of disagreement, as does the exclusion of outliers.

How should group agreement be measured over many scenarios?

The impacts of different ways of defining group agreement for any one scenario are of interest in their own right. However as a method of measuring **overall** group agreement, the simple proportion of scenarios for which there is panel agreement gives results that can be difficult to interpret. Firstly, even with random responses, some level of agreement will occur, and the real interest lies in how much better the observed agreement is than this. The problem is that the underlying level of chance agreement will depend on how the responses are distributed. If most participants think either that most scenarios are appropriate or that most are inappropriate, there will be higher levels of chance agreement for each scenario than there would be if responses were more evenly spread across the scale.

Secondly, reducing the group response for each scenario to 'disagree' involves very substantial loss

of information: the nature and extent of the disagreement is not captured. Thirdly, there is the issue of what is meant by overall agreement. Participants may agree about the **relative** positions of each scenario on the scale of appropriateness (e.g. that scenario A is more appropriate than scenario B), while disagreeing about the **absolute** positions (e.g. that scenario A is highly appropriate). This section describes methods of measuring agreement that seek to address these problems.

Naylor and colleagues (1990) also looked at the impact of changing definitions of consensus on more sophisticated indicators of agreement. In one analysis they computed the probability that any two randomly drawn participants would agree on a particular rating (*Pa* in *Table 25*). For agreement on the exact scale point the probability was 29% overall (with variation depending on the type of scenario). For agreement within a 3-point range the level was 65%. However, this comparison was not corrected for agreement due to chance. Using the kappa statistic as a measure of agreement, the level of agreement for the exact scale point was 16% and the level of agreement for a 3-point range was 43%.

TABLE 25 *Overall agreement among NGT participants (%) considering acceptable delay for coronary revascularisation (Naylor* et al, *1990)*

	Exact rating	Category (3-point range)
Unanimity	0	0.11
3/4 majority	0.01	0.59
Pa[*]	0.29	0.65
kappa	0.16	0.43
ICC	0.51	–

[*] Pa: *probability that any two of 16 participants would agree on a particular scenario*

Naylor and colleagues (1990) went on to calculate intra-class correlation coefficients (ICCs) based on the original 9-point scale ratings. This method uses data from the whole distribution of responses for each scenario rather the agree–disagree dichotomy. (An alternative suggested by Cohen (1968) is the weighted kappa.) The values were generally rather higher than the unweighted kappas using the wider-category 3-point scales: for example, the overall ICC was 0.51 as against a wide-category kappa of 0.43. However, when responses were tightly clustered, the ICC values were close to the kappa values for exact agreement on the 9-point scales of 0.12 to 0.18.

James and colleagues (1984) examined agreement under the assumption that response distributions were biased rather than subject to random error. Priem and colleagues (1995) proposed a group consensus score based on variability in rankings, computing the summed difference between individual and group rankings. That analysis, however, did not look at the extent of agreement between individuals; it indicated simply complete agreement or disagreement, but not degrees of agreement. In order to include this in the analysis, Naylor and colleagues calculated the ICC (an alternative would have been to compute a weighted kappa). This resulted in 51% agreement. However, this analysis was in some ways problematic as a measure of agreement. The problem is evident when different scenarios are compared. For patients with more severe unstable angina, the ICC was similar to the kappa for an exact match of scale points (0.13 and 0.12 respectively). However, if categories (based on 3-point ranges of the scale) were considered, the kappa was 0.70. This was because 88% of ratings for severe unstable angina were clustered in the 1 to 4 range. Thus while there is little variability in the categorical ratings, there is disagreement about the exact scale point, making the ICC small.

Apart from the kappa statistic and ICC analyses which Naylor and colleagues used, other methods exist for calculating the amount of agreement. Such methods expand on the kappa statistic to allow for the extent of agreement to be taken into account (Cohen, 1968), and also to examine agreement when it is presumed that response distributions are biased rather than random (James *et al,* 1984). Priem and colleagues (1995) analysed the degree of consensus by looking at the variability in rankings. They computed the summed difference between individual and group rankings to produce a group consensus score.

It is not surprising that there is no agreed standard as to how to measure consensus (Kozlowski & Hattrup, 1992; Shrout, 1993). It may be necessary in the analysis of consensus to distinguish between models which examine inter-rater reliability and those which look at inter-rater agreement. Kozlowski and Hattrup (1992) distinguished between reliability (the proportional consistency of variance among raters) and agreement (which looks at the extent to which raters make essentially the same rating). They questioned the appropriateness of the ICC, a measure of reliability, as a measure of

agreement. As this is a measure of reliability it can lead to the type of problem seen above where a low ICC was obtained on some ratings because there was low variability between ratings. The opposite is also possible, that is high reliability may be found when there is little agreement among raters, if the ratings are different but proportional. This is illustrated by the two examples of ratings of ten items on a 9-point scale by two raters which are shown in the box below.

In example 1 reliability is high. The ratings are proportional between raters. However, the two sets of ratings are obviously not in agreement. In example 2 reliability is lower. There is little variability among ratings. However, there is more agreement on the ratings than in the first example.

Further complexity is encountered in measuring consensus when comparing the amount of agreement between two groups (for example, comparing the amount of consensus in a mixed panel with a panel composed of all one speciality). There is disagreement about whether indices of agreement within each group should be computed and then compared across groups, or whether some form of within and between analysis which examines individual variability within and between groups should be used (Yammarino & Markham, 1992).

Summary

Simple proportions of the number of scenarios for which there is group agreement and disagreement are difficult to interpret and provide no information on the extent or distribution of divergent views. There is, however, no agreed standard method for determining and communicating the level of group consensus. Whichever method is used, it is important to avoid confusing the level of reliability (association) of the participants' views with the level of agreement.

Implications for clinical guideline development

Implicit versus explicit methods

An implicit approach to aggregating individual judgements may be adequate for establishing broad policy guidelines but more explicit methods are needed to develop detailed, specific guidelines.

Definitions of agreement

The more demanding the definition of agreement, the more anodyne the consensus statement will be. If the requirement is too demanding, either no statements will qualify or those that do will be of little interest.

Weighting participants' judgements

Differential weighting of individual participants' views produces unreliable results unless there is a clear empirical basis for calculating the weights.

Outliers

The exclusion of individuals with extreme views (outliers) can have a marked effect on the content of guidelines.

Aggregation methods

There is no agreement as to the best method of mathematical aggregation. Whichever method is used, the report should include an indication of the distribution or dispersal of participants' judgements and not just the measure of central tendency. In general, the median and the interquartile range are more robust than the mean and standard deviation.

Examples of the lack of correlation between reliability and agreement

Example 1										
Item no. ...	1	2	3	4	5	6	7	8	9	10
Rater 1	1	3	5	1	3	5	1	3	5	1
Rater 2	5	7	9	5	7	9	5	7	9	5
Example 2										
Item no. ...	1	2	3	4	5	6	7	8	9	10
Rater 1	1	2	2	3	1	3	2	3	1	4
Rater 2	2	1	2	1	3	2	3	3	3	3

Chapter 9

Implications and recommendations for further research

Good practice in clinical guideline development

A considerable amount of research has been carried out on consensus development methods, but many aspects have not been investigated sufficiently. For the time being at least, advice on these aspects has therefore to be based on the user's own common sense and the experience of those who have used or participated in these methods. To avoid confusion, the extent to which research support for any guidance on good practice is indicated as follows:

- A = clear research evidence
- B = limited supporting research evidence
- C = experienced common-sense judgement.

B and C should not be regarded as necessarily unsatisfactory as some aspects of the conduct of consensus development methods are not amenable to scientific study but can be adequately justified on the basis of experience.

The three principal formal methods (the Delphi method, the NGT, and the consensus development conference) have been described in chapter 1. The aim of this guide is to highlight issues of general concern when using consensus development methods for developing clinical guidelines. Many of these issues arise whichever method is used. For some issues, advice specific to particular methods is necessary and is provided.

Defining the task or constructing the questions

- Suitable topics for guideline development are those in which there is a mismatch between clinical practice and available research evidence. Conversely, if a clear and prima facie appropriate consensus already exists on the topic (as evidenced by little variation in clinical practice and clear research evidence), there is little scope for a formal consensus method to improve clinical practice. [C]
- It is necessary to decide whether to focus on ways of managing a specific condition or on ways of using a specific intervention (investigation or treatment). The former has two advantages: (1) it reflects the clinician's role (that is deciding how best to investigate or treat a patient); (2) it ensures that participants consider alternatives when judging the appropriate use of any single intervention. [C]
- If the topic is an intervention, it is necessary to decide whether the intervention is to be rated alone or alongside specified alternatives. [C]
- Once the topic has been clearly defined, it is necessary to identify all the relevant cues that will affect participants' judgements of good practice. Cues that are likely to be relevant for many topics include the patient's age, the severity of the condition and the extent of any co-morbidity. Cues can be identified from a review of the literature and from an initial survey of participants. While the cues to be considered are best drawn from published literature, there are benefits from involving participants in the process as they are likely to contribute other, often important, information. Cues suggested by participants, for which research evidence is lacking, should be given the benefit of the doubt and included. This will help maintain group cohesion and cooperation. [C]
- A balance is required between comprehensiveness and brevity. Participants' estimations of probabilities are sensitive to the level of detail provided and the number of options offered. In general, a more comprehensive approach is recommended as it stimulates participants to consider issues that they may otherwise overlook. There is, however, no advantage in seeking to include all possible scenarios. It is preferable to focus on those that are common or clinically important. It is better for a group to examine a small number of scenarios in detail than a large number superficially. [B]
- The context in which the decision is to be made (such general cues as whether participants should assume unlimited healthcare resources or the reality of restricted resources) is important and should be made explicit. Differences of context at least partly explain why similarly

composed groups in different countries sometimes arrive at different group judgements on the same topics when considering the same research evidence. [B]

Selecting the participants

- Group decisions will to some extent reflect the profession or specialty of the participants. Consensus-based guidelines should therefore be interpreted in the context of the composition of the group. [A]
- Having decided on the broad composition of a group in terms of professional and specialty mix, there is little evidence as to whether or not the selection of individuals has much effect on the output. There may, however, be a greater probability of the output being widely accepted if the group is seen to be credible by the target audience and to reflect a relevant range of opinion. Selection should also be seen to be unbiased. [C]
- If the aim is simply to identify areas of agreement, groups should be homogeneous in composition. If, in contrast, the aim is to identify and explore areas of uncertainty, a heterogeneous group is appropriate. For clinical guidelines, it is not just a matter of agreement but correctness, for which heterogeneous groups may have an advantage. [B]
- For NGTs and consensus conference panels, groups should be large enough to produce reasonable reliability in combined individual ratings, but not so large as to be unmanageable as a group. Practical issues in organising and managing face-to-face group interaction suggests that the optimal number is around ten participants, a size that provides adequate reliability. [A]
- Because the status of participants is known to affect their contribution to and influence within a group, efforts should be made to mitigate the effects of status, such as by ensuring individual judgements remain confidential and by skilful chairing or facilitation. [C]
- When face-to-face interaction does not occur (Delphi method), the larger the group size the better in terms of ownership and hence acceptance of the results. Again, diminishing returns suggest that above group sizes of about 15 participants, improvements in reliability are quite small. [B]

Providing information

- As the research evidence on any particular topic is of variable methodological quality, it is important that people skilled in research methods are involved in conducting any literature review. This will increase the likelihood of judgements being based on valid and appropriate information. [C]
- Grading the methodological quality of studies using a reliable method (standard, validated checklists) may mitigate the biases of the reviewers somewhat, but may not eliminate them. [B]
- Information should be presented in a synthesised way which is easy to read and understand, and which brings out the dimensions on which judgements should be based, for example study design. Tables and charts which summarise the available research evidence are preferable to the original research papers. Information which is novel is more likely to be remembered and used. [C]
- Differences in prior beliefs and attitudes to risk are likely to lead to participants interpreting the same information differently. [A]

Information for NGTs and conferences

- Research-based information should be provided to participants for five reasons: (1) it has an impact; (2) if all members of the group have access to this same information it is more likely to be discussed within the group; (3) providing a literature review to group members before discussion may enhance the perception that the task is research-based, which will encourage members to be more reliant on information; (4) if participants come to the discussion with opinions that are, at least to some extent, based on a reading of relevant research, the information exchanged may be more likely to reflect the research evidence; (5) providing a common starting point may foster group cohesion. [C]
- Group members should be encouraged to bring the review, and any notes they have made on it, to the group sessions as a memory aid. [B]

Information for Delphi studies

- Feedback given to participants in Delphi studies should include the arguments deployed as well as simple frequency distributions of individuals' ratings. [B]

Choosing a method of structuring the interaction

- Formal methods generally perform better than informal ones. [B]
- Different methods may give rise to different outcomes, with none being clearly more appropriate than another. [B]
- As the ways the methods are used in practice varies, the actual procedures used should be carefully documented and described in any reports. [C]

- In NGTs a comfortable environment for meetings is likely to be conducive to discussion. A good facilitator will enhance consensus development. The facilitator should guide discussion and should avoid stating his or her views at an early stage. [C]
- If an initial consensus exists, the group should be made aware that it faces the danger of shifting to a more extreme view than any individual may be comfortable with. [A]
- With the NGT and the Delphi method, two or more rating rounds will generally produce some convergence of individual judgements, though it is unclear whether this increases the reliability of the group decision. [A]

Choosing a method of combining individual judgements

- An implicit approach or an approach based on simple voting may be sufficient for establishing broad policy guidelines but some more sensitive form of mathematical aggregation is needed to develop specific guidelines. [C]
- There is no agreement as to the best method of mathematical aggregation. [B]
- Where the frequency distribution of the individual judgements is not obviously multimodal, an appropriate measure of central tendency should be used. Generally the median is preferred to the mean as it is more robust to the effect of outliers. [A]
- An indication of the distribution or spread of participants' judgements should also be reported. This ensures that the audience will have some indication of the extent of consensus. The inter-quartile range is more robust than the standard deviation. [A]
- If results are presented in terms of the proportion agreeing, the exclusion of outliers can have a marked effect. Also, the more demanding the definition of agreement, the more anodyne the outcome will be, so that if the requirement is too demanding, either no statements will qualify or those that do will be of little interest. This approach is not recommended. [A]
- Differential weighting of individual participants' views should be avoided unless there is some empirical basis for it, such as data on past performance in similar tasks. This is unlikely to be available in the context of clinical guidelines. [B]

Future research agenda

Some questions are amenable to research because they can be answered. There are also many questions to which there are no answers in terms of what is correct or best. For these questions, however, it may be possible to find out what effect a particular factor has on the process or outcome of consensus development.

As will be apparent from this review, research is needed in many areas of consensus development. Realistically, it will be possible to fund only a limited amount of research on consensus development in the healthcare sector, though relevant studies in other sectors will no doubt continue in parallel. We have, therefore, focused on the five areas of uncertainty which appear to offer the greatest potential benefit for improving our understanding and application of consensus methods.

What impact does the framing or presentation of the question have on individual judgement?

For example, to what extent are ratings of appropriateness of one intervention on its own different from ratings of the same intervention in the context of other treatments? Would making other treatments salient affect the ratings?

In what form and how inclusive should scenarios be?

Research is needed to compare the ratings obtained using theoretical questions (such as, which age groups of patients would you treat?) with case-based scenarios (in which scenarios of patients of different ages are rated). Does judgement vary by the type of scenario? Are some types of scenario easier to judge than others? Do clinicians draw on different types of information in rating different types of scenarios? Are case-based judgements richer in information than theoretical scenarios? How do clinicians react to making these judgements?

How does the extent of heterogeneity of a group affect the process and outcome?

The effects of homogeneous and heterogeneous groups in terms of specialty background needs further investigation. Comparisons have been made between mixed groups and surgeons, but homogeneous groups of non-surgeons (including primary care doctors) have not been investigated. Also, the balance of specialty groups within a panel could be investigated – for example, heterogeneous panels with an equal number of members of each specialty versus panels which include unequal numbers. Analysis of both the processes and outcome within the group is desirable. Do minority views get properly aired? If so, are they eliminated during discussion or by the process of aggregation?

What effect does research-based information have on individual and on group judgements? Does the effect depend on the amount of information or how it is presented?

This could be examined by studying group interactions or by examining the effect of prompts from the facilitator regarding the information. Studies could compare groups which have or have not been provided with literature reviews, or which are given evidence presented in different ways (written, oral presentation, video), or which are provided with different amounts of information. Such studies would aim to determine whether the information is exchanged within groups and whether the exchange of information affects the outcome. What is the optimal amount of information that should be presented? What techniques can be used to make the information presented more accessible?

What effect does the method of feedback of participants' views have on group judgement?

Studies of different ways of providing feedback on group views to individuals are needed. These could examine the effect on outcome, inter-individual understanding and discussion within the group. Using a Delphi method, what difference does it make whether or not the feedback is attributed to a named individual?

References

Abelson RP (1976). Script processing in attitude formation and decision making. In: Carroll JS, Payne JW, editors. Cognition and social behaviour. Hillsdale (NJ): Erlbaum.

Abrams D, Hogg MA (1990). Social identification, self-categorization, and social influence. *Eur Rev Soc Psychol*;**1**:195–228.

Agency for Health Care Policy and Research (1995). AHCPR Clinical Practice Guideline Program. Report to Congress. US Department of Health and Human Services, AHCPR.

Anderson LE, Balzer WK (1991). The effects of timing of leaders opinions on problem solving groups: a field experiment. *Group Organ Stud*;**16**:86–101.

Anson R, Bostrom R, Wynne B (1995). An experiment assessing group support system and facilitator effects on meeting outcomes. *Manage Sci*;**41**:189–208.

Argyle M, Dean J (1965). Eye-contact, distance and affiliation. *Sociometry*;**28**:289–304.

Arrow KJ (1963). Social choice and individual values. New Haven (CT): Yale University Press.

Asch SE (1956). Studies of independence and conformity: a minority of one against a unanimous majority. *Psychol Monogr*,**70 (9, whole no. 416)**.

Baddeley AD (1986). Working memory. Oxford: Oxford University Press.

Bales RF (1970). Personality and interpersonal behaviour. New York: Holt, Rinehart & Winston.

Bales RF, Stodtbeck FL (1951). Phases in group problem-solving. *J Abnorm Soc Psychol*;**46**:485–95.

Ballard DJ, Etchason JA, Hilborne LH, Campion ME, Kamberg CJ, Soloman DH, *et al* (1992). Abdominal aortic aneurism surgery: a literature review and ratings of appropriateness and necessity. Santa Monica (CA): RAND. Report JRA-04.

Bantel KA (1993a). Strategic clarity in banking: role of top management-team demography. *Psychol Rep*;**73**:1187–201.

Bantel KA (1993b). Comprehensiveness of strategic planning: the importance of heterogeneity of a top team. *Psychol Rep*;**73**:35–49.

Baron J (1988). Thinking and deciding. Cambridge: Cambridge University Press.

Bartlett FC (1932). Remembering: a study in experimental and social psychology. Cambridge: Cambridge University Press.

Bayley EW, Richmond T, Noroian EL, Allen LR (1994). A Delphi study on research priorities for trauma nursing. *Am J Crit Care*;**3**:208–16.

Bell DW, Raiffa H, Tversky A, editors (1988). Decision making. Descriptive, normative and prescriptive interactions. Cambridge: Cambridge University Press.

Berner ES, Webster GD, Shugerman AA, Jackson JR, Algina J, Baker AL, *et al* (1994). Performance of four computer based diagnostic systems. *New Engl J Med*;**330**:1792–6.

Bernstein SJ, Laouri M, Hilborne LH, Leape LL, Kahan JP, Park RE, *et al* (1992). Coronary angiography: a literature review and ratings of appropriateness and necessity. Santa Monica (CA): RAND. Report JRA-03.

Bernstein SJ, Hofer TP, Meijler AP, Rigter H (1997). Setting standards for effectiveness: a comparison of expert panels and decision analysis. *Int J Qual Health Care*;**9**:255–64.

Bettman JR, Kakkar P (1977). Effects of information presentation format on consumer information acquisition strategies. *J Consumer Res*;**3**:233–40.

Black NA (1994). Appropriateness of medical care in Europe: a commentary. *Int J Qual Health Care*;**6**:231–2.

BMJ (1988). Consensus conference. Treatment of stroke. *BMJ*;**297**:126–8.

Boje DM, Murnighan K (1982). Group confidence pressures in iterative decisions. *Manage Sci*;**2**:1187–96.

Bond S, Bond J (1982). A Delphi study of clinical nursing research priorities. *J Adv Nurs*;**7**:565–75.

Bottger PC (1984). Expertise and air time as bases of actual and perceived influence in problem-solving groups. *J Appl Psychol*;**69**:214–21.

Brehm SS, Brehm JW (1981). Psychological reactance: a theory of freedom and control. New York: Academic Press.

Brehmer B, Joyce CRB, editors (1988). Human judgement: the SJT view. Amsterdam: Elsevier.

Brightman HJ, Lewis DJ, Verhoeven P (1983). Nominal and interacting groups as Bayesian information processors. *Psychol Rep*;**53**:101–2.

Brook RH, Williams KN (1976). Effect of medical care review on the use of injections. A study of the New Mexico experimental medical care review organization. *Ann Intern Med*;**85**:509–15.

Brook RH, Kosecoff JB, Park E, Chassin MR, Winslow CM, Hampton JR (1988). Diagnosis and treatment of coronary disease: comparison of doctors' attitudes in the USA and the UK. *Lancet*;**ii**:750–3.

Brown WJ, Redman S (1995). Setting targets: a three-stage model for determining priorities for health promotion. *Aust J Publ Health*;**19**:263–9.

Brown RL, Brown RL, Edwards JA, Nutz JF (1992). Variation in a medical faculty's decisions to transfuse. Implications for modifying blood product utilization. *Med Care*;**30**:1083–93.

Buchan MS, Hegge MJ, Stenvig, TE (1991). A tiger by the tail: tackling barriers to differentiated practice. *J Contin Educ Nurs*;**22**:109–12.

Bucher HC, Weinbacher M, Gyr K (1994). Influence of method of reporting study results on decision of physicians to prescribe drugs to lower cholesterol concentration. *BMJ*;**309**:761–4.

Burke PJ (1971). Task and socio-emotional leadership role performance. *Sociometry*;**34**:22–40.

Burleson BR, Levine BJ, Samter W (1984). Decision-making procedure and decision quality. *Hum Commun Res*;**10**:557–74.

Burns TJ, Batavia AI, Smith QW, De Jong G (1990). Primary health care needs of persons with physical disabilities: what are the research and service priorities? *Arch Phys Med Rehabil*;**71**:138–43.

Burnstein E, Vinokur A. (1975). What a person thinks upon learning he has chosen differently from others: nice evidence of the persuasive arguments explanation of choice shifts. *J Exp Soc Psychol*;**11**:412–26.

Can J Cardiol (1994). 2nd Canadian consensus conference on arrhythmias: validation strategies for the treatment of arrhythmias. Montreal, September 18, 1992. *Can J Cardiol*;**10**:323–41.

Can J Cardiol (1995). Canadian consensus conference on coronary thrombolysis – recommendations 1994. *Can J Cardiol*;**11**:587–95.

Can Med Assoc J (1991a). Consensus conference on Lyme disease. *Can Med Assoc J*;**144**:1627–32.

Can Med Assoc J (1991b). Assessing dementia: the Canadian consensus. Organizing committee, Canadian consensus conference on the assessment of dementia. *Can Med Assoc J*;**144**:851–3.

Centre for Reviews and Dissemination (1995). Undertaking systematic reviews of research on effectiveness. York: University of York. CRD Report 4.

Cannon GS, Idol L, West JF (1992). Educating students with mild handicaps in general classrooms: essential teaching practices for general and special educators. *J Learning Disabilities*;**25**:300–17.

Chassin M (1989). How do we decide whether an investigation or procedure is appropriate? In: Hopkins A, editor. Appropriate investigation and treatment in clinical practice. London: Royal College of Physicians. p. 21–9.

Clark RD, Willems EP (1969). Where is the risky shift? *J Pers Soc Psychol*;**13**:215–21.

Clawson VK, Bostrom RP, Anson R (1993). The role of the facilitator in computer supported meetings. *Small Group Res*;**24**:547–65.

Cohen J (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement and partial credit. *Psychol Bull*;**70**:213–20.

Coleman A (1982). Game theory and experimental games. Oxford: Pergamon Press.

Conn Med (1991). Consensus conference. Treatment of early stage breast cancer. National Institutes of Health. *Conn Med*;**55**:101–7.

Coulter I, Adams A, Shekelle P (1995). Impact of varying panel membership on ratings of appropriateness in consensus panels: a comparison of a multi- and single disciplinary panel. *Health Serv Res*;**30**:577–91.

Crepaldi G, Belfiore F, Bosello O, Caviezel F, Contaldo F, Enzi G, *et al* (1991). Italian Consensus Conference – overweight, obesity and health. *Int J Obes*;**15**:81–90.

Dalkey NC (1969). An experimental study of group opinion: the Delphi method. *Futures*;**1**:408–26.

Dalkey NC, Helmer O (1963). An experimental application of the Delphi method to the use of experts. *Manage Sci*;**9**:458–67.

Davis JH (1973). Group decision and social interaction: a theory of social decision schemes. *Psychol Rev*;**80**:97–125.

Davis JH (1980). Group decision and procedural justice. In: Fishbein M, editor. Progress in social psychology. Hillsdale, NJ: Erlbaum.

Davis JH (1992). Some compelling intuitions about group consensus decisions, theoretical and empirical research, and interpersonal aggregation phenomena: selected examples, 1950–1990. *Organ Behav Hum Decis Processes*;**52**:3–38.

Davis JH, Kerr NL, Atkin RS, Holt R, Meek D (1975). The decision processes of 6- and 12-person mock juries assigned unanimous and two thirds majority rules. *J Pers Soc Psychol*;**32**:1–14.

Davis JH, Kameda T, Parks C, Stasson M, Zimmerman S (1989). Some social mechanics of group decision making: the distribution of opinion polling sequence, and implications for consensus. *J Pers Soc Psychol*;**57**:1000–12.

Davis JH, Stasson MF, Parks CD, Hulbert L, Kameda T, Zimmerman SK, Ono K (1993). Quantitative decisions by groups and individuals: voting procedures and monetary awards by mock civil juries. *J Exp Soc Psychol*;**29**:326–46.

Delbecq A, Van de Ven A (1971). A group process model for problem identification and program planning. *J Appl Behav Sci*;**7**:467–92.

Diehl M, Stroebe W (1987). Productivity loss in brainstorming groups: toward the solution of a riddle. *J Pers Soc Psychol*;**53**:497–509.

Deutsch M (1973). The resolution of conflict: constructive and destructive processes. New Haven (CT): Yale University Press.

Deutsch M, Gerard HB (1955). A study of normative and informational social influences upon individual judgment. *J Abnorm Soc Psychol*;**51**:629–36.

Driskell JE, Salas E (1991). Group decision making under stress. *J Appl Psychol*;**76**:473–8.

Duffield C (1993). The Delphi technique: a comparison of results obtained using two expert panels. *Int J Nurs Stud*;**30**:227–37.

Durand-Zaleski I, Bonnet F, Rochant H, Bierling P, Lemaire F (1992). Usefulness of consensus conferences: the case of albumin. *Lancet*:**340**:1388–90.

Elstein AS, Bordage G (1988). Psychology of clinical reasoning. In: Dowie J, Elstein A, editors. Professional judgement. Cambridge: Cambridge University Press.

Erffmeyer RC, Lane IM (1984). Quality and acceptance of an evaluative task: the effects of four group decision-making formats. *Group Organ Stud*;**9**:509–29.

Evans JStBT, Harries C, Dennis I, Dean J (1995). General practitioners' tacit and stated policies in the prescription of lipid lowering agents. *Br J Gen Pract*;**45**:15–18.

Fahey T, Griffiths S, Peters TJ (1995). Evidence based purchasing: understanding results of clinical trials and systematic review. *BMJ*;**311**:1056–60.

Felson DT, Anderson JJ, Boers M, Bombardier C, Chernoff M, Fried B, *et al* (1993). The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. The Committee on Outcome Measures in Rheumatoid Arthritis Clinical Trials. *Arthritis Rheum*;**36**:729–40.

Felsenthal DS, Fuchs E (1976). Experimental evaluation of five designs of redundant organizational systems. *Admin Sci Q*;**21**:474–87.

Ferguson JH (1996). The NIH consensus development program. The evolution of guidelines. *Int J Technol Assess Health Care*;**12**:460–74.

Ferrell WR (1985). Combining individual judgments. In: Wright G, editor. Behavioral decision making. New York: Plenum Press.

Fiedler FE (1967). A theory of leadership effectiveness. New York: McGraw-Hill.

Fink A, Kosecoff J, Chassin M, Brook RH (1984). Consensus methods: characteristics and guidelines for use. *Am J Publ Health*;**74**:979–83.

Fischer GW (1981). When oracles fail – a comparison of four procedures for aggregating subjective probability forecasts. *Organ Behav Hum Performance*;**28**:96–110.

Fischhoff B, Slovic P, Lichtenstein S (1978). Fault trees: sensitivity of estimated failure probabilities to problem representation. *J Exp Psychol Hum Percept Performance*;**4**:330–44.

Fiske ST, Taylor SE (1984). Social Cognition. Reading (MA): Addison-Wesley.

Fletcher SW (1997). Whither scientific deliberation in health policy recommendations? *New Engl J Med*;**336**:1180–3.

Flores BE, White EM (1989). Subjective vs objective combining of forecasts: an experiment. *J Forecasting*;**8**:331–41.

Flowers ML (1977). A laboratory test of some implications of Janis's groupthink hypothesis. *J Pers Soc Psychol*;**35**:888–96.

Fraser GM, Pilpel D, Hollis S, Kosecoff J, Brook RH (1993). Indications for cholecystectomy: the results of a consensus panel approach. *Qual Assur Health Care*;**5**:75–80.

Freedman JL, Fraser SC (1966). The bases of social power. In: Cartwright D, editor. Studies in social power. Ann Arbor (MI): University of Michigan Press.

French JRP Jr, Raven BH (1959). The bases of social power. In: Cartwright D, editor. Studies in social power. Ann Arbor (MI): University of Michigan Press.

Genest C, Zidek JV (1986). Combining probability distributions: a critique and an annotated bibliography. *Stat Sci*;**1**:114–48.

George JF, Dennis AR, Nunamaker JF (1992). An experimental investigation of facilitation in an EMS decision room. *Group Decis Negotiation*;**1**:57–70.

Gigerenzer G, Hell W, Blank H (1988). Presentation and content: the use of base rates as a continuous variable. *J Exp Psychol Hum Percept Performance*;**14**:513–25.

Goplerud EN, Walfish S, Broskowski A (1985). Weathering the cuts: a Delphi survey on surviving cutbacks in community mental health. *Community Ment Health J*;**21**:14–27.

Gowan JA, McNichols CW (1993). The effects of alternative forms of knowledge representation on decision-making consensus. *Int J Man Machine Stud*;**38**:489–507.

Granger CWJ (1989). Combining forecasts – 20 years later. *J Forecasting*;**8**:167–73.

Green SG, Taber TD (1980). The effects of three decision schemes on decision group process. *Organ Behav Hum Performance*;**25**:97–106.

Grimshaw JM, Russell IT (1993). Effect of clinical guidelines on medical practice: a systematic review of rigorous evaluations. *Lancet*;**342**:1317–22.

Grol R, van Beurden W, Binkhorst T, Toemen T (1991). Patient education in family practice: the consensus reached by patients, doctors and experts. *Fam Pract*;**8**:133–9.

Gustafson DH, Shukla RK, Delbecq A, Walstre GW (1973). A comparative study of differences in subjective likelihood estimates made by individuals, interacting groups, Delphi groups, and nominal groups. *Organ Behav Hum Performance*;**9**:280–91.

Guzzo RA, Dickson MW (1996). Teams in organizations: recent research on performance and effectiveness. *Annu Rev Psychol*;**47**:307–38.

Haas S (1993). European consensus statement on the prevention of venous thromboembolism. European Consensus Conference, Windsor, UK, November, 1991. *Blood Coagul Fibrinolysis*;**4 Suppl 1**: S5–S8; discussion S9–S10.

Haight BK, Bahr RT (1992). Setting an agenda for clinical nursing research in long-term care. *Clin Nurs Res*;**1**:144–57.

Hakim S, Weinblatt J (1993). The Delphi process as a tool for decision making: the case of vocational training of people with handicaps. *Eval Program Planning*;**16**:25–38.

Hall J, Watson WH (1970). The effects of a normative interaction on group decision-making performance. *Hum Relations*;**23**:299–317.

Hammond KR, Brehmer B (1973). Quasi-rationality and distrust: implications for international conflict. In: Rappoport L, Summers DA, editors. Human judgment and social interaction. New York: Holt, Rinehart, & Winston.

Harrington JM (1994). Research priorities in occupational medicine: a survey of United Kingdom medical opinion by the Delphi technique. *Occup Environ Med*;**51**:289–94.

Hastie R, Penrod SD, Pennington N (1983). Inside the jury. Cambridge (MA): Harvard University Press.

Hegedus DM, Rasmussen RV (1986). Task effectiveness and interaction process of a modified nominal group technique in solving an evaluation problem. *J Manage*;**12**:545–60.

Heider F (1958). The psychology of interpersonal relations. New York: Wiley.

Herbert TT, Yost EB (1979). A comparison of decision quality under nominal and interacting consensus group formats: the case of the structured problem. *Decis Sci*;**10**:358–70.

Higgins ET, Bargh JA (1987). Social cognition and social perception. *Annu Rev Psychol*;**38**:369–425.

Hilborne LH, Leape LL, Kahan JP, Park RE, Kamberg CJ, Brook RH (1992). Percutaneous transluminal coronary angioplasty: a literature review and ratings of appropriateness and necessity. Santa Monica (CA): RAND. Report JRA-01.

Hiss RG, Greenfield S (1996). Forum Three: changes in the U.S. health care system that would facilitate improved care for non-insulin-dependent diabetes mellitus. *Ann Intern Med*;**124 (1 Pt 2)**:180–3.

Hogarth M (1978). A note on aggregating opinions. *Organ Behav Hum Performance*;**21**:40–6.

Hornsby JS, Smith BN, Gupta JND (1994). The impact of decision making methodology on job evaluation outcomes: a look at three consensus approaches. *Group Organ Manage*;**19**:112–28.

Hovland C, Janis I, Kelley HH (1953). Communication and persuasion. New Haven (CT): Yale University Press.

Huber GP, Delbecq A (1972). Guidelines for combining the judgments of individual members in decision conferences. *Acad Manage J*;**15**:161–84.

Hunter DJW, McKee CM, Sanderson CFB, Black NA (1994). Appropriate indications for prostatectomy in the UK – results of a consensus panel. *J Epidemiol Community Health*;**4**:58–64.

Imamura K, Gair R, McKee M, Black N (1997). Appropriateness of total hip replacement in the United Kingdom. *World Hosp Health Serv*;**32**:10–14.

Innami I (1994). The quality of group decisions, group verbal-behaviour, and intervention. *Organ Behav Hum Decis Processes*;**60**:409–30.

Isen AM, Means B (1983). The influence of positive affect on decision making strategy. *Soc Cogn*;**2**:18–31.

Isenberg DJ (1986). Group polarization: a critical review and meta-analysis. *J Pers Soc Psychol*;**50**:1141–51.

Jacoby I (1988). Evidence and consensus. *JAMA*;**259**:3039.

Jackson SE (1992). Team composition in organizational settings: issues in managing an increasingly diverse work force. In: Worchel S, Wood W, Simpson J, editors. Group process and productivity. Newbury Park (CA): Sage. p. 138–73.

James LR, Demaree RG, Wolf G (1984). Estimating within-group interrater reliability with and without response bias. *J Appl Psychol*;**69**:85–98.

Janis I (1982). Groupthink. 2nd ed. Boston: Houghton-Mifflin.

Jarboe SC (1988). A comparison of input-output, process-output, and input-process-output models of small group problem-solving. *Commun Monogr*;**55**:121–42.

Jarvenpaa SL (1990). Graphic displays in decision making – the visual salience effect. *J Behav Decis Making*;**3**:247–62.

Jehn KA (1995). A multimethod examination of the benefits and detriments of intragroup conflict. *Admin Sci Q*;**40**:256–82.

Jones J, Hunter D (1995). Consensus methods for medical and health services research. *BMJ*;**311**:377–80.

Jones J, Sanderson C, Black N (1992). What will happen to the quality of care with fewer junior doctors? A Delphi study of consultant physicians' views. *J Roy Coll Phys Lond*;**26**:36–40.

Kahan JP, Park RE, Leape LL, Bernstein SJ, Hilborne LH, Parker L, *et al* (1996). Variations by specialty in physician ratings of the appropriateness and necessity of indications for procedures. *Med Care*;**34**:512–23.

Kameda T, Sugimori S (1993). Psychological entrapment in group decision making: an assigned decision rule and a groupthink perspective. *J Pers Soc Psychol*;**65**:282–92.

Kaplan MF (1987). The influencing process in group decision making. In: Hendrick C, editor. Review of personality and social psychology. No. 8. Group processes. Newbury Park (CA): Sage.

Kaplan MF, Miller CE (1987). Group decision making and normative vs. informational influence: effects of type of issue and assigned decision rule. *J Pers Soc Psychol*;**53**:306–13.

Karau SJ, Kelly JR (1992). The effects of time scarcity and time abundance on group performance quality and interaction process. *J Exp Soc Psychol*;**28**:542–71.

Karma P, Palva T, Kouvalainen K, Karja J, Makela PH, Prinssi VP, *et al* (1987). Finnish approach to the treatment of acute otitis media. Report of the Finnish Consensus Conference. *Ann Otol Rhinol Laryngol*;**129 Suppl**:1–19.

Kastein MR, Jacobs M, Van der Hell RH, Luttik K, Touw-Otten FWMM (1993). Delphi, the issue of reliability: a qualitative Delphi study in primary health care in the Netherlands. *Technol Forecasting Soc Change*;**44**:315–23.

Kelley HH (1967). Attribution theory in social psychology. *Nebraska Symp Motivation*;**15**:192–240.

Kirchler E, Davis JH (1986). The influence of member status differences and task type on group consensus and member position change. *J Pers Soc Psychol*;**51**:83–91.

Koehler JJ (1993). The influence of prior beliefs on scientific judgements of evidence quality. *Organ Behav Hum Decis Processes*;**56**:28–55.

Korman AF (1966). Consideration, initiating structures, and organizational criteria: a review. *Personnel Psychol*;**19**:349–62.

Kors JA, Sittig AC, van Bemmel JH (1990). The Delphi method to validate diagnostic knowledge in computerized ECG interpretation. *Methods Inf Med*;**29**:44–50.

Kosecoff J, Kanouse DE, Rogers WH, McCloskey L, Winslow CM, Brook RH (1987). Effects of the National Institutes of Health consensus development program on physician practice. *JAMA*;**258**:2708–13.

Kovess V (1995). The French Consensus Conference on long-term therapeutic strategies for schizophrenic psychoses. *Soc Psychiatry Psychiatr Epidemiol*;**30**:49–52.

Kozlowski SJ, Hattrup K (1992). A disagreement about within-group agreement: disentangling issues of consistency versus consensus. *J Appl Psychol*;**77**:161–7.

Lancet (1997). The prostate question, unanswered still (editorial). *Lancet*;**349**:443.

Larach MG, Localio AR, Allen GC, Denborough MA, Ellis FR, Gronert GA, *et al* (1994). A clinical grading scale to predict malignant hyperthermia susceptibility. *Anesthesiology*;**80**:771–9.

Larreche JC, Moinpour R (1983). Managerial judgment in marketing: the concept of expertise. *J Marketing Res*;**20**:110–21.

Laupacis A, Sackett DL, Roberts R (1988). An assessment of clinically useful measures of the consequences of treatment. *New Engl J Med*;**318**:1728–33.

Leape LL, Freshour MA, Yntema D, Hsiao W (1992a). Small group judgment methods for determining resource based relative values. *Med Care*;**30 11 Suppl**:NS28–NS39.

Leape LL, Hilborne LH, Kahan JP, Stason WB, Park RE, Kamberg CJ, *et al* (1992b). Coronary artery bypass graft: a literature review and ratings of appropriateness and necessity. Santa Monica (CA): RAND. RAND/AMCC Report JRA-02.

Leape LL, Park RE, Kahan JP, Brook RH (1992c). Group judgments of appropriateness: the effect of panel composition. *Qual Assur Health Care*;**4**:151–9.

Lee PP, Kamberg CJ, Hilborne LH, Massanari RM, Kahan JP, Park RE, *et al* (1993). Cataract surgery: a review of the literature regarding efficacy and risks. Santa Monica, CA: RAND. RAND/AMCC Report JRA-06.

Lethbridge DJ, McClurg V, Henrikson M, Wall G (1993). Validation of the nursing diagnosis of ineffective breastfeeding. *J Obstet Gynecol Neonatal Nurs*;**22**:57–63.

Levine JM, Moreland RL (1990). Progress in small group research. *Annu Rev Psychol*;**41**:585–634.

Lichtenstein S, Slovic P, Fischoff B, Layman M, Coombs B (1978). Judged frequency of lethal events. *J Exp Psychol*;**4**:58–64.

Loftus EF (1979). Eyewitness testimony. Cambridge (MA): Harvard University Press.

Lomas J (1991). Words without action? The production, dissemination, and impact of consensus recommendations. *Annu Rev Publ Health*;**12**:41–65.

Lomas J, Pickard L, Mohide A (1987). Patient versus clinician item generation for quality-of-life measures. The case of language-disabled adults. *Med Care*;**25**:764–9.

Lomas J, Anderson G, Enkin M, Vayda E, Roberts R, Mackinnon B (1988). The role of evidence in the consensus process: results from a Canadian consensus exercise. *JAMA*;**259**:3001–5.

Lord CG, Ross L, Lepper MR (1979). Biased assimilation and attitude polarization: the effects of prior theories on subsequently considered evidence. *J Pers Soc Psychol*;**37**:2098–109.

Maass A, Clark RD III (1984). Hidden impact of minorities: fifteen years of minority influence research. *Psychol Bull*;**95**:428–50.

McClellan M, Brook RH (1992). Appropriateness of care – a comparison of global and outcome methods to set standards. *Med Care*;**30**:565–86.

McDonald CJ, Hui SL, Smith DM, Tierney WM, Cohen SJ, Weinberger M, *et al* (1984). Reminders to physicians from an introspective computer medical record. *Ann Intern Med*;**10**:130–8.

McGrath JE (1978). Small group research. *Am Behav Sci*;**21**:651–74.

McGrath JE (1984). Groups: interaction and performance. Englewood Cliffs (NJ): Prentice-Hall.

McKee M, Black N (1993). Junior doctors' work at night: what is done and how much is appropriate? *J Publ Health Med*;**15**:16–24.

McKee M, Priest P, Ginzler M, Black N (1991). How representative are members of expert panels? *Qual Assur Health Care*;**3**:89–94.

McNeil BJ, Pauker SG (1984). Decision analysis for public health: principles and illustrations. *Annu Rev Publ Health*;**5**:135–61.

McNeil BJ, Pauker SG, Sox HC, Tversky A (1982). On the elicitation of preferences for alternative therapies. *New Engl J Med*;**306**:1259–62.

McNeil BJ, Pauker SG, Tversky A (1988). On the framing of medical decisions. In: Bell D, Raiffa H, Tversky A, editors. Decision making: descriptive, normative and prescriptive interactions. Cambridge: Cambridge University Press. p. 562–8.

Maier NRF, McRay P (1972). Increasing innovation in change situations through leadership skills. *Psychol Rep*;**31**:343–54.

Maier NRF, Sashkin M (1971). Specific leadership behaviours that promote problem solving. *Personnel Psychol*;**24**:35–44.

Markus H, Zajonc RB (1985). The cognitive perspective in social psychology. In: Lindzey G, Aronson E, editors. Handbook of social psychology. 3rd ed. Vol 1. New York: Random House.

Mann T (1996). Clinical guidelines. Using clinical guidelines to improve patient care in the NHS. London: Department of Health.

Matcher DB, Goldstein LB, McCrory DC, Oddone EZ, Jansen DA, Hilborne LH, *et al* (1992). Carotid endarterectomy: a literature review and ratings of appropriateness and necessity. Santa Monica (CA): RAND. Report JRA-05.

Maznevski ML (1994). Understanding our differences: performance in decision-making groups with diverse members. *Hum Relations*;**47**:531–52.

Med J Aust (1994). The management of hypertension: a consensus statement. Australian Consensus Conference 1993. *Med J Aust*;**160 Suppl**:S1–16.

Meehl PE (1954). Clinical versus statistical prediction: a theoretical analysis and review of the literature. Minneapolis: University of Minnesota Press.

Mehrabian A (1972). Nonverbal communication. Chicago: Aldine-Atherton.

Merrick NJ, Fink A, Park RE, Brook RH, Kosecoff J, Chassin MR, *et al* (1987). Derivation of clinical indications for carotid endarterectomy by an expert panel. *Am J Publ Health*;**77**:187–90.

Miner FC (1979). A comparative analysis of three diverse group decision-making approaches. *Acad Manage J*;**22**:81–93.

Mobily PR, Herr KA, Kelley LS (1993). Cognitive-behavioral techniques to reduce pain: a validation study. *Int J Nurs Stud*;**30**:537–48.

Moscovici S (1976). Social influence and social change. New York: Academic Press.

Moscovici S (1985). Social influence and conformity. In: Lindzey G, Aronson E, editors. Handbook of social psychology. 3rd ed. Vol 2. New York: Random House.

Murray AI (1989). Top management group heterogeneity and firm performance. *Strategic Manage J*;**10**:125–41.

Nagao DH, Davis JH (1980). Some implications of temporal drift in social parameters. *J Exp Soc Psychol*;**16**:479–96.

Naylor CD, Basinski A, Baigrie RS, Goldman BS, Lomas J (1990). Placing patients in the queue for coronary revascularization: evidence for practice variations from an expert panel process. *Am J Publ Health*;**80**:1246–52.

Nemeth CJ (1992). Minority dissent as a stimulant to group performance In: Worchel S, Wood W, Simpson J, editors. Group process and productivity. Newbury Park (CA): Sage.

Nemiroff RM, King DD (1975). Group decision-making as influenced by consensus and self orientation. *Hum Relations*;**28**:1–21.

Nemiroff RM, Pasmore WA, Ford DL (1976). The effects of two normative structural interventions on established and ad hoc groups: implications for improving decision making effectiveness. *Decis Sci*;**7**:841–55.

NIH Consensus Development Panel on Epilepsy (1990). National Institutes of Health Consensus Conference. Surgery for epilepsy. *JAMA*;**264**:729–33.

NIH Consensus Development Panel on Melanoma (1992a). National Institutes of Health Consensus Conference. Diagnosis and treatment of early melanoma. *JAMA*;**268**:1314–9.

NIH Consensus Development Panel on Depression (1992b). National Institutes of Health Consensus Conference. Diagnosis and treatment of depression in late life. *JAMA*;**268**:1018–24.

NIH Consensus Development Panel on Impotence (1993). National Institutes of Health Consensus Conference. Impotence. *JAMA*;**270**:83–90.

NIH Consensus Development Panel on Optimal Calcium Intake (1994). National Institutes of Health Consensus Conference. Optimal calcium intake. *JAMA*;**272**:1942–8.

NIH Consensus Development Panel on Total Hip Replacement (1995a). National Institutes of Health Consensus Conference. Total hip replacement. *JAMA*;**273**:1950–6.

NIH Consensus Development Panel on Ovarian Cancer (1995b). NIH consensus conference. Ovarian cancer. Screening, treatment, and follow-up. *JAMA*;**273**:491–7.

Nisbett R, Ross L (1980). Human inference: strategies and shortcomings of social judgment. New Jersey: Prentice-Hall.

Nisbett RE, Wilson TD (1977). Telling more than we can know: verbal reports on mental processes. *Psychol Rev*;**84**:231–59.

Oddone EZ, Samsa G, Matchar DB (1994). Global judgments versus decision-model-facilitated judgments: are experts internally consistent? *Med Decis Making*;**14**:19–26.

Orne MT (1969). Demand characteristics and the concept of quasi-controls. In: Rosenthal R, Rosnow R, editors. Artifact in behaviour research. New York: Academic Press.

Parente FJ, Anderson-Parente JK (1987). Delphi inquiry systems. In: Wright G, Ayton P, editors. Judgmental forecasting. Chichester: Wiley. p. 129–56.

Parente FJ, Anderson JK, Myers P, O'Brien T (1984). An examination of factors contributing to Delphi accuracy. *J Forecasting*;**3**:173–82.

Park RE, Fink A, Brook RH, Chassin MR, Kahn KL, Merrick NJ, *et al* (1986). Physician ratings of appropriate indications for six medical and surgical procedures. Santa Monica (CA): RAND. R-3280-CWF/HF/PMT/RJW.

Park RE, Fink A, Brook RH, Chassin MR, Kahn KL, Merrick NJ, *et al* (1989). Physician ratings of appropriate indications for three procedures: theoretical indications vs indications used in practice. *Am J Publ Health*;**79**:445–7.

Passannante MR, Gallaghe, CT, Reichman LB (1994). Preventive therapy for contacts of multidrug-resistant tuberculosis. A Delphi survey. *Chest*;**106**:431–4.

Pavitt C (1993). What (little) we know about formal group discussion procedures. A review of relevant research. *Small Group Res*;**24**:217–35.

Payne JW, Bettman JR, Johnson EJ (1992). Behavioral decision research: a constructive processing perspective. *Annu Rev Psychol*;**43**:87–131.

Pearson SD, Margolis CZ, Davis C, Schreier LK, Sokol HN, Gottlieb LK (1995). Is consensus reproducible? A study of an algorithmic guidelines development process. *Med Care*;**33**:643–60.

Petty RE, Cacioppo JT (1977). Forewarning, cognitive responding and resistance to persuasion. *J Pers Soc Psychol*;**35**:645–55.

Petty RE, Cacioppo JT (1981). Attitudes and persuasion: classic and contemporary approaches. Dubuque (IA): William C. Brown.

Pill J (1971). The Delphi method: substance, context, a critique and the annotated bibliography. *Socioecon Planning Sci*;**5**:57–71.

Plumridge RJ (1981). Forecast of the future of hospital pharmacy in Australia. *Am J Hosp Pharm*;**38**:1469–72.

Podsakoff PM, Schriesheim CA (1985). Field studies of French and Raven's bases of power: critique, reanalysis and suggestions for future research. *Psychol Bull*;**97**:387–411.

Priem RL, Price KH (1991). Process and outcome expectations for the dialectical inquiry, devil's advocacy, and consensus techniques of strategic decision making. *Group Organ Stud*;**16**:206–25.

Priem RL, Harrison DA, Muir NK (1995). Structured conflict and consensus outcomes in group decision making. *J Manage*;**21**:691–710.

Pruitt DG (1981). Negotiation behaviour. New York: Academic Press.

Reagan-Cirincione P, Rohrbaugh J (1992). Decision conferencing: a unique approach to the behavioral aggregation of expert judgment. In: Wright G, Bolger F, editors. Expertise and decision support. New York: Plenum Press. p. 181–201.

Redelmeier DA, Shafir E (1995). Medical decision making in situations that offer multiple alternatives. *JAMA*;**273**:302–5.

Redelmeier DA, Koehler DJ, Liberman V, Tversky A (1995). Probability judgment in medicine: discounting unspecified possibilities. *Med Decis Making*;**15**:227–30.

Richardson FMacD (1972). Peer review of medical care. *Med Care*;**10**:29–39.

Rinaldi RC, Steindler EM, Wilford BB, Goodwin D (1988). Clarification and standardization of substance abuse terminology. *JAMA*;**259**:555–7.

Rogelberg SG, Barnes Farrell JL, Lowe CA (1992). The stepladder technique: an alternative group structure. *J Appl Psychol*;**77**:730–7.

Rohrbaugh J (1979). Improving the quality of group judgment: social judgment analysis and the Delphi technique. *Organ Behav Hum Performance*;**24**:73–92.

Rohrbaugh J (1981). Improving the quality of group judgment: social judgment analysis and the nominal group technique. *Organ Behav Hum Performance*;**28**:272–88.

Roth PL (1994). Group approaches to the Schmidt-Hunter global estimation procedure. *Organ Behav Hum Decis Processes*;**59**:428–51.

Rowe G (1992). Perspectives on expertise in the aggregation of judgments. In: Wright G, Bolger F, editors. Expertise and decision support. New York: Plenum Press. p. 155–80.

Rowe G, Wright G, Bolger F (1991). Delphi: a reevaluation of research and theory. *Technol Forecasting Soc Change*;**39**:235–51.

Rubin JZ, Brown BR (1975). The social psychology of bargaining and negotiation. New York: Academic Press.

Russo JE (1977). The value of unit price information. *J Marketing Res*;**14**:193–201.

Schank RC, Abelson RP (1977). Scripts, plans, goals and understanding. Hillsdale (NJ): Lawrence Erlbaum Associates Inc.

Schittekatte M (1996). Facilitating information exchange in small decision-making groups. *Eur J Soc Psychol*;**26**:537–56.

Schneider DJ, Hastorf AH, Ellsworth PC (1979). Person perception. 2nd ed. Reading (MA): Addison-Wesley.

Schweiger DM, Sandberg WR, Ragan JW (1986). Group approaches for improving strategic decision making. *Acad Manage J*;**29**:51–71.

Schweiger D, Sandberg W, Rechner P (1989). Experiential effects of dialectical inquiry, devil's advocacy and consensus approaches to strategic decision making. *Acad Manage J*;**32**:745–72.

Schwenk C, Cosier R (1993). Effects of consensus and devil's advocacy on strategic decision-making. *J Appl Soc Psychol*;**23**:126–39.

Scott EA, Black N (1991a). When does consensus exist in expert panels? *J Publ Health Med*;**13**:35–9.

Scott EA, Black N (1991b). Appropriateness of cholecystectomy in the United Kingdom – a consensus panel approach. *Gut*;**32**:1066–70.

Shanteau J (1992). How much information does an expert use? Is it relevant? *Acta Psychol*;**81**:75–86.

Shanteau J, Nagy GF (1979). Probability of acceptance in dating choice. *J Pers Soc Psychol*;**37**:522–33.

Shaw ME (1981). Group dynamics. The psychology of small group behaviour. 3rd ed. New York: McGraw-Hill.

Shekelle PG, Adams AJ, Chassin MR, Hurwitz FL, Phillips RB, Brook RH (1991). The appropriateness of spinal manipulation of low-back pain: project overview and literature review. Santa Monica (CA): RAND. Publication No. R-4025/1-CCR/FCER.

Sherif M (1937). An experimental approach to the study of attitudes. *Sociometry*;**1**:90–8.

Sherif M, Hovland C (1961). Social judgment. New Haven (CT): Yale University Press.

Sherif M, Sherif CW (1979). Research on intergroup relations. In: Austin WG, Worchel S, editors. The social psychology of intergroup relations. Pacific Grove (CA): Brooks/Cole.

Shrout PE (1993). Analyzing consensus in personality judgments: a variance components approach. *J Pers*;**61**:769–88.

Silverstein MD, Ballard DJ (1998). Expert panel assessment of appropriateness of abdominal aortic aneurysm surgery: global judgment versus probability estimates. *J Health Serv Res Policy* (in press).

Slovic P, Fischhoff B, Lichtenstein S (1988). Response mode, framing, and information-processing effects in risk assessment. In: Bell DE, Raiffa H, Tversky A, editors. Decision making. Descriptive, normative and prescriptive interactions. Cambridge: Cambridge University Press. p. 152–66.

Sniezek JA (1990). A comparison of techniques for judgmental forecasting by groups with common information. *Group Organ Stud*;**15**:5–19.

Sommer R, Olsen H (1980). The soft classroom. *Environ Behav*;**5**:3–16.

Soon A, O'Connor M (1991). The effect of group interaction processes on performance in time series extrapolation. *Int J Forecasting*;**7**:141–9.

Souder WE (1977). Effectiveness of nominal and interacting group decision processes for integrating R&D and marketing. *Manage Sci*;**23**:595–605.

Stasser G (1992). Pooling of unshared information during group discussion. In: Worchel S, Wood W, Simpson J, editors. Group process and productivity. Newbury Park (CA): Sage. p. 48–67.

Stasser G, Titus W (1987). Effects of information load and percentage of shared information on the dissemination of unshared information during group discussion. *J Pers Soc Psychol*;**53**:81–93.

Stasser G, Taylor LA, Hanna C (1989). Information sampling in structured and unstructured discussion of three- and six-person groups. *J Pers Soc Psychol*;**57**:67–78.

Steiner ID (1972). Group process and productivity. New York: Academic Press.

Stephenson BY, Michaelsen LK, Franklin SG (1982). An empirical test of the nominal group technique in state solar energy planning. *Group Organ Stud*;**7**:320–34.

Stewart J, Kendall E, Cooke A (1994). Citizens' juries. London: IPPR.

Strauss G, Chassin M, Lock J (1995). Can experts agree when to hospitalize adolescents? *J Am Acad Child Adolesc Psychiatry*;**34**:418–24.

Stocking B (1985). First consensus development conference in United Kingdom: on coronary artery bypass grafting. 1. Views of audience, panel, and speakers. *BMJ*;**292**:713–6.

Strube MJ, Garcia JE (1981). A meta-analytical investigation of Fiedler's contingency model of leadership effectiveness. *Psychol Bull*;**90**:307–21.

Tajfel H (1982). Social psychology of intergroup relations. *Annu Rev Pyschol*;**33**:1–39.

Tajfel H, Turner JC (1986). The social identity theory of intergroup behaviour. In: Worchel S, Austin WG, editors. Psychology of intergroup relations. 2nd ed. Chicago: Nelson-Hall.

Taylor SE, Crocker J (1981) Schematic bases of social information processing. In: Social cognition: the Ontario symposium. Higgins ET, Herman CP, Zanna MP, editors. Hillside (NJ): Erlbaum.

Tepper S, De Jong G, Wilderson D, Brannon R (1995). Criteria for selection of a payment method for inpatient medical rehabilitation. *Arch Phys Med Rehabil*;**76**:349–54.

Tournageau R, Rasinski KA, Bradburn N, D'Andrade R (1989). Belief accessibility and context effects in attitude measurement. *J Exp Soc Psychol*;**25**:401–21.

Tversky A, Kahneman D (1974). Judgment under uncertainty: heuristics and biases. *Science*;**185**:1124–31.

Tversky A, Kahneman D (1981). The framing of decisions and the psychology of choice. *Science*;**211**:453–8.

Tversky A, Koeler DJ (1994). Support theory: a nonextensional representation of subjective probability. *Psychol Rev*;**101**:547–67.

Van de Ven AH, Delbecq AL (1974). The effectiveness of nominal, Delphi, and interacting group decision making processes. *Acad Manage J*;**17**:605–21.

Vinokur A, Burnstein E (1978a). Novel argumentation and attitude change: the case of polarization following group discussion. *Eur J Soc Psychol*;**11**:127–48.

Vinokur A, Burnstein E (1978b). The depolarization of attitudes in groups. *J Pers Soc Psychol*;**36**:872–85.

Vinokur A, Burnstein E, Sechrest L, Wortman PM (1985). Group decision making by experts: field study of panels evaluating medical technologies. *J Pers Soc Psychol*;**49**:70–84.

Weiner B (1974). Achievement motivation and attribution theory. Morristown (NJ): General Learning Press.

White SE, Dittrich JE, Lang JR (1980). The effects of group decision-making process and problem-situation complexity on implementation attempts. *Admin Sci Q*;**25**:428–39.

Whitehead M, Lobo RA (1988). Consensus conference: progestagen use in postmenopausal women. *Lancet*;**2**:1243–4.

Whitney JC, Smith RA (1983). Effects of group cohesiveness on attitude polarization and the acquisition of knowledge in a strategic planning context. *J Marketing Res*;**20**:167–76.

Wiersema MF, Bantel KA (1992). Top management team demography and corporate strategic change. *Acad Manage J*;**35**:91–121.

Wilke HAM, Meertens RW (1994). Group performance. London: Routledge.

Williams S, Taormina RJ (1993). Unanimous versus majority influences on group polarization in business decision making. *J Soc Psychol*;**133**:199–205.

Worchel S, Shackelford SL (1991). Groups under stress: the influence of group structure and environment on process and performance. *Pers Soc Psychol Bull*;**17**:640–7.

Wortman PM, Vinokur A, Sechrest L (1988). Do consensus conferences work? A process evaluation of the NIH consensus development program. *J Health Polit Policy Law*;**13**:469–98.

Woudenberg F (1991). An evaluation of Delphi. *Technol Forecasting Soc Change*;**40**:131–50.

Yammarino FJ, Markham SE (1992). On the application of within and between analysis: are absence and affect really group-based phenomena? *J Appl Psychol*;**77**:168–76.

Zadinsky JK, Boettcher JH (1992). Preventability of infant mortality in a rural community. *Nurs Res*;**41**:223–7.

Zajonc RB (1965). Social facilitation. *Science*;**149**:269–74.

Appendix 1

Impact of clinical guidelines

The impact of guidelines can be assessed in terms of their influence on practice (an intermediate outcome) or whether they affect patients' health (the final outcome). It is important to recognise that their impact concerns not only the process of production but also the process of dissemination.

Grimshaw and Russell (1993) conducted a systematic review of studies which had examined the effect of clinical guidelines on practice. They defined clinical guidelines as 'systematically developed statements to assist practitioner decisions about appropriate health care for specific clinical circumstances'. Only some of the guidelines reviewed were derived from consensus groups and, although many of the studies reviewed used guidelines produced by a group, few of them had been produced by formal consensus development methods.

Grimshaw and Russell (1993) reviewed 59 papers and concluded that all but four detected significant improvements in the process of care following the introduction of guidelines (i.e. the process changed in line with the guidelines). Only 11 studies looked at the effects on patient health, of which nine reported significant improvements. However, the size of the improvements, for both the process of care and patient outcomes, varied considerably between studies.

The studies differed in the type of guideline development strategy, the way guidelines were disseminated and the way they were implemented. Some guidelines were generated through consensus development conferences with results published in national journals with no particular implementation strategies, others were developed by local users, disseminated through group discussion or other face-to-face education, and implemented by providing specific reminders to doctors at the time of patient consultation. To illustrate the range of studies, a small selection are described below.

McDonald and colleagues (1984) developed a computer program incorporating some 1491 rules to generate 751 different reminder messages about good clinical practice. These rules were developed by a committee of three general medicine faculty members and other consulting subspecialists. The reminders covered a wide range of conditions and types of care. Thus while the rules were the output of a group, the type of group and the type of output differed considerably from consensus groups which focus on a single issue. Within one hospital some doctors were provided with computer reminders of these guidelines at the time of consultation, whereas others were not. Thus the intervention strategy was very timely and direct. The reminded group were more likely to implement treatment in accordance with the guidelines.

Brook and Williams (1976) examined the effects of locally produced guidelines combined with financial incentives (no reimbursement for inappropriate use) on the rate of injections given to the Medicaid population. The number of injections fell significantly, although inappropriate use was not entirely eliminated.

Two studies which investigated guidelines produced through consensus development conferences used different implementation techniques and obtained different results. Kosekoff and colleagues (1987) examined whether the quality of care increased as a result of guidelines produced by the consensus conferences. Using patient records, they examined care in ten hospitals for 24 months preceding and 13 to 24 months after each consensus development conference. No particular implementation strategy was used; results of the conferences were largely disseminated through publications. Kosekoff and colleagues (1987) concluded that taken as a whole the four consensus conferences had no effect on practice.

Durand-Zaleski and colleagues (1992) looked at the effect of consensus development conference guidelines on the use of albumin in one French hospital. They found a substantial reduction in the use of albumin, in accordance with guidelines, following their dissemination. However, the dissemination of these guidelines was quite different from that in the study by Kosekoff and colleagues (1987). The hospital disseminated the guidelines, held meetings with doctors, monitored the use of albumin, and provided feedback to the prescribers.

Thus the range of production and dissemination techniques of guidelines is very wide. Grimshaw and Russell (1993) suggested that when the development strategy was internal, when dissemination involved a specific educational intervention, and when implementation involved patient-specific reminders at the time of consultation, such strategies were more effective in producing change towards the guidelines. In contrast, guidelines produced externally or nationally, disseminated in journals, and implemented with only a general reminder produced lower rates of compliance.

Lomas (1991) has reviewed ten evaluations of the impact of consensus-based recommendations on practice. Six studies found no impact, two found a minor impact and two found a major impact. He noted that three of the four studies which showed an impact were from Europe. This suggests that guidelines produced by consensus development methods may have less impact than guidelines produced through other means. However, this finding may be confounded by the method of dissemination of the guidelines – consensus recommendations often rely simply on publication for dissemination.

Appendix 2

Details of studies comparing consensus development methods

Subjective likelihood (probability) estimation and forecasting tasks

Seven studies have used either probability estimation or forecasting tasks. Of these, two compared the NGT, the Delphi method, staticised groups and informal groups (Fischer *et al,* 1981; Gustafson *et al,* 1973) and a third compared the NGT, the Delphi method and staticised groups (Boje & Murnighan, 1982). All three studies investigated probability estimation – though Boje & Murnighan (1982) also examined answers to almanac questions (see below). Gustafson and colleagues (1973) found that the NGT performed better than any of the other methods, that is, groups using the NGT were more accurate in the estimates of probability than groups using other methods. However, Fischer and colleagues (1981) found no differences in performance between the different methods. Boje and Murnighan (1982) also found no differences, but they did find that over repeated trials of estimation, with both the Delphi method and the NGT group, estimates became less accurate although the actual differences between groups were fairly small. However, in both the study by Gustafson and colleagues (1973) and that by Fischer and colleagues the Delphi method used involved only one round of feedback.

In the study by Gustafson and colleagues (1973) the members were all in the same room. Boje and Murnighan (1982) also had subjects in the same room but behind partitions. They used three rounds and fed back not only the probability estimate but also the reasons for the estimate. In the NGTs in the studies by Gustafson and colleagues (1973) and Fischer and colleagues there was open discussion after giving estimates whereas in the study by Boje and Murnighan (1982) a researcher led a round-robin format. The study by Fischer and colleagues (1981) study involved ten trials at estimating a probability, and after each trial, subjects were given feedback on their accuracy. The feedback may have swamped any differences that existed between the groups.

Two other studies compared the NGT with other methods. Brightman and colleagues (1983) compared the NGT with informal methods on a probability estimation task. They found that the NGT was significantly closer to the standard than the informal groups. Soon and O'Connor (1991) compared the NGT, staticised groups, informal groups and a group in which one member makes an individual estimate which is then presented to the group for discussion. The task was time-series forecasts in which participants had to make six-point estimates. They had two levels of problem difficulty, easy and difficult. The methods used were poorly explained. Looking at the percentage of error compared with the actual time-series data, they found that for easy problems there was little difference between consensus methods. However, for difficult problems the group in which an individual made an initial estimate performed better than the other groups, which all performed similarly. In both of these studies there is no mention of a facilitator. In the study by Brightman and colleagues (1983), participants were said to be trained in the NGT and they followed a round-robin format. In Soon and O'Connor's (1991) study, groups were provided with documentation explaining the method they were to use. It may be that the groups 'ran' themselves.

Another study using five forecasting problems compared the Delphi method, staticised groups, informal groups and the 'best member' technique (in which the group discusses the task face-to-face and then selects the member who they think performs best) (Sneizek, 1990). This member's initial judgement becomes the group's judgement. All members of all groups gave individual pre-group forecasts. Participants in the Delphi method were in the same room and had median estimates fed back. The criteria for stopping were either the same median results three times in a row or three of the five members agreeing on the forecast.

No significant differences were found between the methods for two simple problems. For the three difficult tasks, no method outperformed the actual best forecast (the opinion of the individual member who was closest to the correct forecast) and no method was superior to any other. On the problem for which individual judgement was

unbiased (error was random), both the staticised group and the actual best forecast were significantly better than all other groups and the 'best member' groups were significantly worse. Thus the Delphi method performed no better than other methods. For one task the Delphi method was worse than the staticised group but better than the 'best member' groups.

Another study (Larreche & Moinpour, 1983) compared the Delphi method, staticised groups, informal groups, and two methods of selecting the 'best member'. The best member was selected either by self-rated confidence in their judgement or by an external measure of expertise. The task was to forecast a market share. All individuals in all groups made individual pre-group judgements. Members of the Delphi groups were in the same room. Average, lowest and highest estimates were fed back as well as reasons why the estimate was lower or higher than the average. There were three iterations. The Delphi method produced more accurate estimates than either staticised or informal groups. Self-rated experts were no better than any of the above methods but experts identified through external means were.

In summary, for probability estimation and forecasting tasks, direct comparisons of the NGT and the Delphi method show the NGT to be better in one instance and no better in two others. In comparisons of the NGT with informal groups, two studies show the NGT to be better and two show no difference. In studies comparing the NGT with staticised groups, one shows the NGT to be better and three show no difference. In studies comparing the Delphi method with staticised groups, four show no difference (except for one task on which the Delphi method was worse), and one shows it to be better. Three studies have compared the Delphi method with interacting groups: one shows the Delphi method to be better and two show no difference.

Ranking tasks

A common ranking task is the NASA moon problem (or a variation on it). Participants rank the value of a number of items in terms of their usefulness for survival on the moon. The group rankings are then compared with expert rankings. The closer the group ranking comes to the expert ranking the better the group is said to have performed.

One study of this type compared the NGT and informal groups (Herbert & Yost, 1979). Another compared the NGT, informal groups and 'consensus' groups (given a set of instructions as to how to reach agreement, such as to avoid arguing for your own ranking and changing your mind to avoid conflict) (Nemiroff *et al,* 1976). A third study compared these three methods as well as the Delphi method in which rankings and rationales were fed back over five iterative rounds (Erffmeyer & Lane, 1984).

In the comparisons of the NGT with informal groups, one study showed that the NGT was better (Herbert & Yost, 1979), one showed no difference (Nemiroff *et al,* 1976) and the other showed that NGT groups were worse (Erffmeyer & Lane, 1984). Both of the comparisons of the NGT with 'consensus groups' found that the consensus groups outperformed the NGT (Erffmeyer & Lane, 1984; Nemiroff *et al,* 1976). In the study in which the Delphi method was examined, it was found to be superior to all other methods.

In all three studies (Erffmeyer & Lane, 1984; Herbert & Yost, 1979; Nemiroff *et al,* 1976) all members of all groups ranked the items before interaction. In the study in which the NGT was found to be better than an informal group (Herbert & Yost, 1979), both methods appear to have included leaders, though their status is unclear. The NGT followed the round-robin format. In the study in which the NGT was found to be worse than either informal or 'consensus' groups, the NGT groups were led by a researcher, the round-robin format was used and individuals provided both ranks and rationales. In the study in which the NGT was not as good as 'consensus' groups, but no worse than informal groups, there seem to have been no facilitators. Instead the group members ran the group themselves.

Burleson and colleagues (1984) compared staticised groups, informal groups and the Delphi method (although they called it an NGT), using the NASA moon problem. Members of the Delphi group were in the same room and feedback was through the researcher. Both ranks and rationales were provided. All groups made pre-group rankings. Informal groups produced better decisions than either the Delphi group or the staticised group, and there was no difference between the latter two.

A further study looked at three different methods of aggregating participants' views: consensus (ranks had to be agreed by all members), majority vote, and a nominal voting scheme analogous to that used in the NGT (individuals ranked items independently and these were mathematically

combined) (Green & Taber, 1980). In all groups, individual rankings for three tasks were presented to the groups for discussion. Participants were asked to rank items in the way they thought a reference group would. For example, for 'values of young people', participants ranked the values as they thought young people would. There were no accuracy measures. Instead they looked at questionnaire measures of group process (individual measures).

The study involved 76 participants but the number of groups is unclear. Each group, or set of groups, participated in each aggregation method for a different question. The order of questions remained unchanged but the order of the aggregation methods used was varied. Perceived participation varied depending on the aggregation method used ($F_{(2,146)} = 19.27$, $p < 0.001$): participants felt the most participation with consensus, followed by majority vote and nominal vote (means 3.62, 3.45 and 3.11, respectively; $p = 0.05$ for all comparisons). The nominal vote resulted in less negative socio-emotional feelings than either the consensus or the majority vote (means 1.76, 2.23 and 2.12, respectively). There were no differences in satisfaction with the decision. There tended to be less emergence of leaders in groups using nominal voting.

Idea generation

With idea generation tasks, groups are often asked to generate ideas about some problem. The criteria for judging performance is usually the amount and/or quality of the ideas generated. Four studies have compared different methods for idea generation.

Van de Ven and Delbecq (1974) used defining the job of student dormitory councillors as their task. They compared the NGT, the Delphi method and informal groups, with 20 groups per method. This study is notable for sampling different types of people who have an interest in the issue (student residents, student housing administrators, faculty staff, and academic administrators). Heterogeneous panels were drawn from the different groups through stratified random sampling. Graduate students and professional planners, who were chosen for their skills in conducting meetings, were used as leaders for the NGTs and the informal groups. They were randomly assigned to the two methods and trained to conduct the particular meeting type. The Delphi was conducted by post with two rounds. Van de Ven and Delbecq (1974) used as their measure of quality the number of ideas generated and the satisfaction with the process. In terms of both the amount of ideas and a combined measure of amount and satisfaction the NGT and the Delphi method were both better than informal groups and no different from each other.

White and colleagues (1980) studied the generation of ideas for dealing with common management problems among nurse supervisors. They compared the NGT, informal groups and a structured discussion technique which followed explicit steps for identifying facts, objectives and causes of the problem, generating and choosing alternatives, and identifying actions necessary for implementation. One feature was that discussion was controlled to avoid premature closure on a single option. The methods were crossed with the type of task (easy, medium, complex) in a repeated measures design so that each of three groups of nurses used each method for a different problem. All groups had trained facilitators with a fairly specific brief. The measure of the success of the method was based on a self-reported measure of the number of attempts made to implement the solutions over the following 1 month. This was an individual rather than a group-level measure. Implementation rates were higher for the NGT than for the structured discussion groups for easy problems, and higher for the NGT than for informal groups for easy and complex problems. For medium and complex problems structured discussion groups had higher rates of implementation than informal groups.

Jarboe (1988) compared the NGT with a structured discussion method, similar to that used by White and colleagues (1980). They used two idea generation tasks. Participants had to generate ideas about what could be done to prevent teenagers starting to smoke cigarettes and to drink alcohol, but in one task legal actions could be included whereas in the other they could not. Groups were trained in the procedures and a researcher served as facilitator. The NGT followed the standard format. The NGT produced significantly more ideas than the structured discussion.

Souder (1977) compared the NGT and informal groups for a task to produce a list of development guidelines. He used largely qualitative comparisons and had very few groups (which were also crossed with three leadership styles for the NGT groups). He examined the density of network structure and questionnaire measures of satisfaction and integration, along with other observations of the groups. The NGT groups with leaders high

ratings for structuring and consideration achieved the best communication structures and greatest satisfaction and integration. The NGT groups with leaders with a low rating for structuring and a high rating for consideration had good communication and satisfaction measures. The NGT groups with leaders with low ratings on both aspects achieved less dense interaction networks and less satisfaction and integration, and were similar to the informal groups in these respects. However, these findings were based on only one or two NGT groups.

General knowledge or almanac questions

Two studies have investigated methods using general knowledge questions. Dalkey (1969) reported on two studies conducted at RAND, which investigated the effectiveness of the Delphi method in comparison with informal groups. One study compared an informal group with a Delphi group in the answering of 20 questions. Accuracy was higher for the Delphi groups on 13 questions and higher for the informal groups on seven questions. For the second study, interactive discussion was engaged in between a second and third Delphi round, making it into more like an NGT than a Delphi method. The results were unclear but it seemed that the introduction of a discussion led to more questions being answered accurately.

Felsenthal and Fuchs (1976) compared five methods for estimating the salary of a doctor: informal groups, sequential groups (individuals solved the problem and passed the solution on to the next individual in line), a staticised group (individuals solve the problem, discuss the solution with others and then modify their own solution), the Delphi method (three rounds with participants in different rooms and only estimates were fed back) and a group which they described as mixed but which we describe as an NGT (individuals make individual estimates, discuss these, and then make individual estimates again). Groups had either three or six members. There were 20–48 groups per method. Groups were not instructed that they must reach consensus so there was not necessarily a group solution. The measure of quality was the probability of a correct answer (p of correct answer with no agreement + p of correct answer given some level of agreement).

This was an individual level of analysis based on the probability that some individual (or individuals) in the group gave the correct answer rather than, for example, comparing a mean group score to the actual response. They used the z test for comparing probabilities which allows for tests only between two groups. What they seem to have done is to take the mean of groups which have similar probabilities and compare this with the means of different groups with similar probabilities. Thus in comparing three-member groups, the Delphi method was grouped with informal and mixed groups because there was little difference in the probabilities among these groups and compared with individual and sequential groups. The combined probability of the first three groups was 0.37 and of the latter two was 0.15. For three-member groups, Delphi, mixed and informal methods were more likely than individual and sequential methods to contain individuals with correct answers. For six-member groups, the Delphi method stood alone at 0.61, a value significantly higher than that for all of the other groups combined at 0.22. Thus for six-member groups, individuals in Delphi groups had a higher probability of obtaining a correct answer than individuals in all other groups.

Other tasks

A variety of other tasks have been used, though some did not have measures of accuracy. Hornsby and colleagues (1994) compared informal groups, NGTs and the Delphi method using seven groups per method. The task was to evaluate four jobs in terms of the amount of compensation deemed appropriate. This study had no measure of quality or accuracy. All individuals made pre-group evaluations of the jobs. Hornsby and colleagues (1994) measured the extent of change from the average individual pre-group estimate to the group estimate: the NGT changed considerably downwards, informal groups changed upwards, and the Delphi groups showed no change. However, it was not possible to judge which was the 'right' outcome. These researchers also used questionnaire-based measures of satisfaction and found significant differences between groups: satisfaction with the NGT was significantly lower than satisfaction with informal groups, and satisfaction with the Delphi method was significantly higher than that with either the NGT or informal groups.

Miner (1979) had groups complete a role-play exercise involving trade offs between productivity and worker satisfaction. Three members acted as workers and one as leader. Miner compared the NGT, the Delphi method (maximum of seven iterations; participants in the same room), and PCL which makes use of leadership skills to

approach an issue constructively and involve members. Both the NGT and PCL had trained students as leaders, though the PCL leaders were drawn from more experienced evening-class students and the NGT and Delphi leaders were drawn from introductory management courses. Quality was measured by productivity, solution acceptance (questionnaire-based) and effectiveness (quality × acceptance). For quality, PCL was slightly, but not significantly, better than the others. For acceptance there were no significant differences between methods. However for effectiveness, PCL was better than both the NGT and the Delphi method, which were similar.

Leape and colleagues (1992a) compared individual first-round ratings, a Delphi method and a modified NGT (RAND method). The participants were practising surgeons and there was only one group per method (19 members in the Delphi group and 11 in the NGT group). The task was to produce estimates of the time and work requirements for a number of medical and surgical services. The standard of comparison used was estimates of these services provided by a national survey. The first-round individual rankings from both groups were closest to the national survey estimates (a difference of 11.0% and 12.6%). The Delphi method diverged further from survey estimates (a difference of 15.2%) and the modified NGT diverged further still (a difference of 21.0%). However, using individual judgement as a standard of comparison is problematic. Although the survey was said to be reliable and to have face validity, this does not necessarily mean that individual judgements are unbiased.

In separate studies, Rohrbaugh (1979; 1981) compared the Delphi method, the NGT and SJA, which provides feedback to participants on their judgement policies (the cues and weights they use to make their judgements). In one study, the Delphi method (two rounds) was compared with SJA. Participants had to develop a policy for predicting the future college performance of students on the basis of personality measures. Groups were heterogeneous on the basis of their initial judgements. For SJA groups, participants were given feedback on their own judgement policies based on their individual pre-group judgements. Group members were encouraged to compare their policies, though the interaction was not structured. Delphi group members were also given their individual judgement policies, but only as ranges and medians of the relevant measures. Individuals then defined their own post-group judgement policy. The quality of the group product was judged by the degree of correlation between a judgement policy and the actual performance of 205 student cases. There were no differences between the judgement policies of the different groups. However, the quality of the individual post-group judgements varied by group: individuals who had taken part in SJA groups performed better than those who were in Delphi groups. SJA participants showed greater improvements over their pre-group decisions and their post-group judgements were significantly better than those of Delphi group members.

A comparison of SJA and the NGT on a similar task (devising a policy for predicting horse-race winners) produced similar results. The quality of the group policies did not differ. The individual judgements of SJA participants improved over their pre-group judgements whereas those of NGT group participants did not. However, there was no difference between the post-group policies of SJA participants and NGT group participants.

HTA panel membership

This report was identified as a priority by the Methodology Panel.

Acute Sector Panel

Chair: Professor John Farndon, University of Bristol †

Professor Senga Bond, University of Newcastle-upon-Tyne †

Professor Ian Cameron, Southeast Thames Regional Health Authority

Ms Lynne Clemence, Mid-Kent Health Care Trust †

Professor Francis Creed, University of Manchester †

Professor Cam Donaldson, University of Aberdeen

Mr John Dunning, Papworth Hospital, Cambridge †

Professor Richard Ellis, St James's University Hospital, Leeds

Mr Leonard Fenwick, Freeman Group of Hospitals, Newcastle-upon-Tyne †

Dr David Field, Leicester Royal Infirmary †

Ms Grace Gibbs, West Middlesex University Hospital NHS Trust †

Dr Neville Goodman, Southmead Hospital Services Trust, Bristol †

Mr Ian Hammond, Bedford & Shires Health & Care NHS Trust

Professor Adrian Harris, Churchill Hospital, Oxford

Professor Robert Hawkins, University of Bristol †

Dr Chris McCall, General Practitioner, Dorset †

Professor Alan McGregor, St Thomas's Hospital, London

Mrs Wilma MacPherson, St Thomas's & Guy's Hospitals, London

Professor Jon Nicoll, University of Sheffield †

Professor John Norman, University of Southampton

Dr John Pounsford, Frenchay Hospital, Bristol †

Professor Gordon Stirrat, St Michael's Hospital, Bristol

Professor Michael Sheppard, Queen Elizabeth Hospital, Birmingham †

Dr William Tarnow-Mordi, University of Dundee

Professor Kenneth Taylor, Hammersmith Hospital, London

Diagnostics and Imaging Panel

Chair: Professor Mike Smith, University of Leeds †

Professor Michael Maisey, Guy's & St Thomas's Hospitals, London *

Professor Andrew Adam, UMDS, London †

Dr Pat Cooke, RDRD, Trent Regional Health Authority

Ms Julia Davison, St Bartholomew's Hospital, London †

Professor Adrian Dixon, University of Cambridge †

Professor MA Ferguson-Smith, University of Cambridge †

Dr Mansel Hacney, University of Manchester

Professor Sean Hilton, St George's Hospital Medical School, London

Mr John Hutton, MEDTAP International Inc., London †

Professor Donald Jeffries, St Bartholomew's Hospital, London †

Dr Andrew Moore, Editor, *Bandolier* †

Professor Chris Price, London Hospital Medical School †

Dr Ian Reynolds, Nottingham Health Authority

Professor Colin Roberts, University of Wales College of Medicine †

Miss Annette Sergeant, Chase Farm Hospital, Enfield

Professor John Stuart, University of Birmingham

Dr Ala Szczepura, University of Warwick †

Mr Stephen Thornton, Cambridge & Huntingdon Health Commission

Dr Gillian Vivian, Royal Cornwall Hospitals Trust †

Dr Jo Walsworth-Bell, South Staffordshire Health Authority †

Dr Greg Warner, General Practitioner, Hampshire †

Methodology Panel

Chair: Professor Martin Buxton, Brunel University †

Professor Anthony Culyer, University of York *

Dr Doug Altman, Institute of Health Sciences, Oxford †

Professor Michael Baum, Royal Marsden Hospital

Professor Nick Black, London School of Hygiene & Tropical Medicine †

Professor Ann Bowling, University College London Medical School †

Dr Rory Collins, University of Oxford

Professor George Davey-Smith, University of Bristol

Dr Vikki Entwistle, University of Aberdeen †

Professor Ray Fitzpatrick, University of Oxford †

Professor Stephen Frankel, University of Bristol

Dr Stephen Harrison, University of Leeds

Mr Philip Hewitson, Leeds FHSA

Professor Richard Lilford, Regional Director, R&D, West Midlands †

Mr Nick Mays, King's Fund, London †

Professor Ian Russell, University of York †

Professor David Sackett, Centre for Evidence Based Medicine, Oxford †

Dr Maurice Slevin, St Bartholomew's Hospital, London

Dr David Spiegelhalter, Institute of Public Health, Cambridge †

Professor Charles Warlow, Western General Hospital, Edinburgh †

* Previous Chair
† Current members

continued

continued

Pharmaceutical Panel

Chair: Professor Tom Walley, University of Liverpool †

Professor Michael Rawlins, University of Newcastle-upon-Tyne*

Dr Colin Bradley, University of Birmingham

Professor Alasdair Breckenridge, RDRD, Northwest Regional Health Authority

Ms Christine Clark, Hope Hospital, Salford †

Mrs Julie Dent, Ealing, Hammersmith & Hounslow Health Authority, London

Mr Barrie Dowdeswell, Royal Victoria Infirmary, Newcastle-upon-Tyne

Dr Desmond Fitzgerald, Mere, Bucklow Hill, Cheshire

Dr Alistair Gray, Health Economics Research Unit, University of Oxford †

Professor Keith Gull, University of Manchester

Dr Keith Jones, Medicines Control Agency

Professor Trevor Jones, ABPI, London †

Ms Sally Knight, Lister Hospital, Stevenage †

Dr Andrew Mortimore, Southampton & SW Hants Health Authority †

Mr Nigel Offen, Essex Rivers Healthcare, Colchester †

Dr John Posnett, University of York

Mrs Marianne Rigge, The College of Health, London †

Mr Simon Robbins, Camden & Islington Health Authority, London †

Dr Frances Rotblat, Medicines Control Agency †

Mrs Katrina Simister, Liverpool Health Authority †

Dr Ross Taylor, University of Aberdeen †

Dr Tim van Zwanenberg, Northern Regional Health Authority

Dr Kent Woods, RDRD, Trent RO, Sheffield †

Population Screening Panel

Chair: Professor Sir John Grimley Evans, Radcliffe Infirmary, Oxford †

Dr Sheila Adam, Department of Health*

Ms Stella Burnside, Altnagelvin Hospitals Trust, Londonderry †

Dr Carol Dezateux, Institute of Child Health, London †

Dr Anne Dixon Brown, NHS Executive, Anglia & Oxford †

Professor Dian Donnai, St Mary's Hospital, Manchester †

Dr Tom Fahey, University of Bristol †

Mrs Gillian Fletcher, National Childbirth Trust †

Professor George Freeman, Charing Cross & Westminster Medical School, London

Dr Mike Gill, Brent & Harrow Health Authority †

Dr JA Muir Gray, RDRD, Anglia & Oxford RO †

Dr Ann Ludbrook, University of Aberdeen †

Professor Alexander Markham, St James's University Hospital, Leeds †

Professor Theresa Marteau, UMDS, London

Dr Ann McPherson, General Practitioner, Oxford †

Professor Catherine Peckham, Institute of Child Health, London

Dr Connie Smith, Parkside NHS Trust, London

Dr Sarah Stewart-Brown, University of Oxford †

Ms Polly Toynbee, Journalist †

Professor Nick Wald, University of London †

Professor Ciaran Woodman, Centre for Cancer Epidemiology, Manchester

Primary and Community Care Panel

Chair: Dr John Tripp, Royal Devon & Exeter Healthcare NHS Trust †

Professor Angela Coulter, King's Fund, London *

Professor Martin Roland, University of Manchester *

Dr Simon Allison, University of Nottingham

Mr Kevin Barton, East London & City Health Authority †

Professor John Bond, University of Newcastle-upon-Tyne †

Ms Judith Brodie, Age Concern, London †

Dr Nicky Cullum, University of York †

Professor Shah Ebrahim, Royal Free Hospital, London

Mr Andrew Farmer, Institute of Health Sciences, Oxford †

Ms Cathy Gritzner, The Patients' Association †

Professor Andrew Haines, RDRD, North Thames Regional Health Authority

Dr Nicholas Hicks, Oxfordshire Health Authority †

Professor Richard Hobbs, University of Birmingham †

Professor Allen Hutchinson, University of Sheffield †

Mr Edward Jones, Rochdale FHSA

Professor Roger Jones, UMDS, London

Mr Lionel Joyce, Chief Executive, Newcastle City Health NHS Trust

Professor Martin Knapp, London School of Economics & Political Science

Professor Karen Luker, University of Liverpool

Professor David Mant, NHS Executive South & West †

Dr Fiona Moss, North Thames British Postgraduate Medical Federation †

Professor Dianne Newham, King's College London

Professor Gillian Parker, University of Leicester †

Dr Robert Peveler, University of Southampton †

Dr Mary Renfrew, University of Oxford

Ms Hilary Scott, Tower Hamlets Healthcare NHS Trust, London †

* Previous Chair
† Current members

National Coordinating Centre for Health Technology Assessment, Advisory Group

Chair: Professor John Gabbay, Wessex Institute for Health Research & Development †

Professor Mike Drummond,
Centre for Health Economics,
University of York †

Ms Lynn Kerridge,
Wessex Institute for Health Research
& Development †

Dr Ruairidh Milne,
Wessex Institute for Health Research
& Development †

Ms Kay Pattison,
Research & Development Directorate,
NHS Executive †

Professor James Raftery,
Health Economics Unit,
University of Birmingham †

Dr Paul Roderick,
Wessex Institute for Health Research
& Development

Professor Ian Russell,
Department of Health, Sciences & Clinical
Evaluation, University of York †

Dr Ken Stein,
Wessex Institute for Health Research
& Development †

Professor Andrew Stevens,
Department of Public Health
& Epidemiology,
University of Birmingham †

† Current members

Health Technology Assessment 1998; Vol. 2: No. 3

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) 1703 595 639 Email: hta@soton.ac.uk
http://www.soton.ac.uk/~hta

ISSN 1366-5278