First aid: Lessons from health economics for economic evaluation in social welfare.

Sarah Byford
Centre for the Economics of Mental Health, Institute of Psychiatry.

Tom Sefton
Centre for Analysis of Social Exclusion and LSE Health and Social Care, London School of Economics and Political Science.

LSE Health and Social Care Discussion Paper Number 4

First published in June 2002 by:
LSE Health and Social Care
The London School of Economics and Political Science
Houghton Street
London
WC2A 2AE

British Library Cataloguing in Publication Data
A catalogue record for this publication is available from the British Library.
ISBN [07530 15714]

Acknowledgements

This paper forms part of a larger research project aimed at developing the infrastructure for economic evaluation in the field of social welfare. The authors would like to thank the Joseph Rowntree Foundation who funded this initiative, Martin Knapp, John Hills and David McDaid, collaborators in the initiative who provided comments on earlier drafts, and the two peer reviewers who provided valuable suggestions for improvement.

Abstract

Beyond health care, the economic evaluation of social welfare programmes is rare but the demand for such evaluation is rising. To encourage greater use of economic evaluation, undoubtedly social welfare researchers need to gain a better understanding of the philosophy behind such evaluation and the methodologies necessary to carry these out in practice. However, there is also a need for economists to better understand the nature of social welfare interventions in order to ensure that the approaches they adopt are appropriate and feasible. This paper explores the principles that guide health economists, the evaluation issues raised by different types of intervention and the techniques used to address them, focusing on the implications for evaluating other social welfare interventions. It argues that there exists a spectrum of complexity in interventions, with many social welfare interventions demonstrating a level of complexity that is less often reached by health care services. This can cause problems for the application of 'traditional' economic evaluation techniques favoured by health economists, such as the measurement of generic outcomes and the randomisation of subjects between interventions. The paper concludes that efforts need to be made to ensure greater scientific validity of evaluations in social welfare research, whilst recognising that the complexity of many interventions may require greater attention to the details of trial design and modifications or additions to the 'conventional' model, such as a greater role for qualitative methods.

About the authors

Sarah Byford is a Senior Lecturer in Health Economics at the Centre for the Economics of Mental Health, Institute of Psychiatry, specialising in the economic evaluation of mental health and social welfare services.
Contact: Centre for the Economics of Mental Health, Box PO24, Institute of Psychiatry, De Crespigny Park, London, SE5 8AF or s.byford@iop.kcl.ac.uk

Tom Sefton is a Research Officer at the ESRC Research Centre for Analysis of Social Exclusion, London School of Economics. He has recently completed a two year project for the Joseph Rowntree Foundation on economic evaluation in the social welfare field.
Contact: Centre for Analysis of Social Exclusion, London School of Economics, Houghton Street, London, WC2A 2AE or t.a.sefton@lse.ac.uk

Contents

Introduction

Beyond healthcare, the economic evaluation of social welfare programmes is rare (McDaid et al, 2002), but the demand for such evaluation, particularly from research funding bodies, is rising. Economic evaluation provides essential evidence to support decisions relating to the allocation of scarce resources and such scarcity is as much a feature of the social welfare sector as the healthcare sector. To encourage greater use of economic evaluation, undoubtedly social welfare researchers need to gain a better understanding of the philosophy behind such evaluation and the methodologies necessary to carry out economic analyses in practice. However, there is also a need for economists to better understand the nature of social welfare interventions in order to ensure that economic methods are appropriate and feasible in practice. This paper looks at the field of health economics in order to draw lessons for the economic evaluation of other areas of social welfare. The aim is to look at the principles that guide health economists, the evaluation issues raised by different types of intervention and the techniques used to address them, focusing on the implications for evaluating other social welfare interventions.

The healthcare sector is a good starting point for the purpose of this paper. First, health economics is a well-developed and established field of research with much literature to draw upon. Quality control systems and guidelines to promote good practice are in place, providing clear criteria for assessing the strengths and weaknesses of individual studies. Secondly, healthcare covers a wide and varied range of interventions and thus confronts the evaluator with many of the issues likely to be faced in other areas of social welfare. The healthcare sector thus provides a good indication of the potential benefits and pitfalls should evaluation in other areas of social welfare develop along similar lines. In particular, it is a test of how far the randomised controlled trial and other quantitatively focused designs, favoured by many economists, could be applied in evaluating other social welfare services. It is in the healthcare sector that the boundaries have been pushed furthest and where these techniques have begun to be used to evaluate social, as well as clinical, interventions.

This is an exploratory paper that sets out the difficulties that may be experienced when undertaking an economic evaluation in a social welfare context, where social welfare encompasses such areas as social care, early intervention schemes, housing, regeneration, community development, work with families and parents and welfare to work. It begins to explore some, but by no means all, of the solutions to these difficulties. More detailed hypotheses for improving the quality and quantity of economic evaluations in this area are explored elsewhere (Sefton et al, 2002; Byford et al, 2002), but further advances will perhaps require empirical research to provide evidence of the success, or otherwise, of economic techniques.

Economic evaluation

The purpose of economic evaluation is to determine the best use of scarce resources. In practice, this means identifying whether a particular intervention is cost-effective relative to other similar

interventions or to other uses for scarce resources. In other words, identifying those services that produce the greatest benefit for a given level of cost or the same level of benefit for less cost. In the healthcare sector, economic evaluation usually involves a systematic attempt to identify, measure and compare all relevant costs and benefits of alternative policies (Drummond et al, 1997). A fairly standard framework is adopted, which links the inputs of different interventions (e.g. use of health services resources) to their outcomes (e.g. morbidity or mortality).

Methods of economic evaluation

All methods of economic evaluation involve the measurement of costs in monetary terms. Outcomes, however, can be presented in a number of different ways. Cost-effectiveness analysis (CEA) is the most commonly adopted approach to economic evaluation in healthcare and involves the measurement of benefits in a single 'natural' or 'condition-specific' outcome measure such as level of blood pressure or depression. The benefits of two or more interventions are combined with their respective costs to provide a measure of cost-effectiveness. Where one intervention is found to dominate (i.e. is more effective and less costly), a formal cost-effectiveness analysis is not essential, since relative cost-effectiveness has already been demonstrated. However, where one intervention is found to produce greater benefit for greater cost, an incremental cost-effectiveness analysis is required. Incremental analysis involves the calculation of the ratio of additional benefits to additional costs produced by one intervention in comparison to another. This cost-effectiveness ratio can then be compared to other interventions employing the same measure of effect, and preference should be given to those with the lowest incremental cost-effectiveness ratio.

The use of natural units of outcome makes CEA easily transferable into social welfare research, where natural units would include such things as jobs created, crimes prevented or measures of social exclusion. CEA does have its weaknesses, however. First, it is impossible to make comparisons across a diverse spectrum of interventions competing for a share of a finite budget. Social welfare services are extremely varied and the aims and outcomes of services provided will differ greatly and often be multiple. Comparisons of cost-effectiveness using natural units can only be made between interventions whose outcomes can be measured on the same scale. Thus, CEA might be used to support funding decisions between two competing schemes for reducing homelessness, but it cannot determine whether the same money would be better spent on a scheme to support older people living in the community.

Secondly, it is difficult to capture all possible effects of an intervention on a single outcome scale. Social services will often influence many areas of an individual's life, but combining costs with multi-dimensional outcomes measured on a number of different scales makes interpretation difficult, particularly if improvements are seen on some scales but not others. To illustrate, a scheme which provides home adaptations for the promotion of independent living for disabled people may have an impact on psychological, social and family functioning, as well as activities of daily living. All these areas can be measured in natural units but a CEA can only be carried out with one outcome scale. Under such circumstances a CEA will often be based on the outcome measure considered to be of greatest importance to the purpose of the evaluation (the primary outcome), requiring some judgement to be made about the relative value of the alternative outcomes of interest.

Where the omission of other outcomes could be misleading, studies may present a range of outcomes (or consequences) alongside the costs, using cost-consequences analysis (CCA). No attempt is made to formally combine costs with benefits and the decision maker is left to form his or her own opinion regarding the relative importance of the alternative outcome scales presented. Although CCA is limited by the inability to rank interventions in terms of cost-effectiveness, in the absence of adequate measures capable of capturing all the consequences of a particular intervention, CCA is likely to be a useful tool for social welfare research. CCA has been used to evaluate complex interventions where outcomes cannot easily be summarised in a single measure (e.g. O'Cathain et al, 1992; Mauskopf et al, 1998; Morrell et al, 2000). The presentation of all costs and consequences can greatly enhance the understanding gained from a CEA and thus CCA should be encouraged even when a primary outcome measure has been selected and a CEA carried out.

An alternative solution to multiple outcomes is to condense them into one generic measure, which is the approach adopted in cost-utility analysis (CUA). As with CEA, a cost-effectiveness ratio can be calculated, but outcomes are measured in terms of utility (level of satisfaction, well-being or quality of life). One example of a utility-based measure is the quality adjusted life year (QALY). The calculation of QALYs involves the use of quality adjustment weights for different health states (e.g. full versus restricted mobility). Once generated, the utility weights are multiplied by the time spent in each state and then summed to provide the number of quality adjusted life years, thereby incorporating the effects of an intervention on the quantity and quality of life. The results are expressed in a cost-utility ratio in terms of the additional cost per QALY gained from undertaking a particular intervention. This provides a common measure of output that allows comparisons to be made between any number of diverse interventions.

The theory behind utility measures is attractive and their importance should not be dismissed, but a number of weaknesses exist that may limit the usefulness of CUA within the field of social welfare, at least in the short-term. Conceptually, the idea of condensing the benefits of a scheme for urban regeneration, for example, into a single outcome measure can be hard to swallow. Such schemes are area-based, rather than being focused on a specific group of individuals; they may influence many people, in many different ways, across many dimensions of life. In addition, utility scales have been criticised for their conceptual foundations, the methodology employed, their lack of sensitivity to change and for ignoring equity considerations (Sen, 1985; Loomes & McKenzie, 1989; Drummond, 1991; Williams & Kind, 1992). Perhaps the main obstacle to the use of CUA in the evaluation of social welfare interventions is the lack of utility scales appropriate to the field. Although a significant quantity of research has been carried out into the development of utility scales for use in health economics, these measures tend to be health focused and may not be broad enough to capture the full impact of social care policies. To increase the usefulness of the utility approach in the social welfare context, appropriate measures must be developed, which will require investment in research.

Cost benefit analysis (CBA) is a method of economic evaluation that is used less often in healthcare than CEA or CUA, although it is more common in some policy areas, such as transport and environment. CBA requires both costs and benefits to be valued in monetary units. It thus becomes possible to directly compare the costs with the benefits of an individual project (i.e. calculate the net benefit) to see which is greater, without the need for a comparator. Like CUA, CBA allows the comparison of any number of diverse interventions and, in addition, it is

possible to make comparisons across different sectors, such as healthcare, education or defence. However, difficulties arise when attempting to value benefits in monetary terms. Methods do exist, such as willingness to pay or observed preferences (Robinson, 1993; Gafni, 1998; Bala et al, 1999), but they are difficult to apply and can be a time consuming and costly addition to an evaluation. Hence CBA in healthcare is relatively rare and the extent of its use in the wider social welfare field is as yet unknown.

One related technique that is and may continue to be more commonly used in the evaluation of social welfare services is the analysis of 'cost-savings'. This is a limited form of CBA that involves the comparison of costs and benefits that are easily converted into monetary units. The costs of an intervention are compared to the savings that are generated through, for example, reductions in crime or school exclusions. 'Outcomes' of this type can be converted into monetary units relatively easily since they involve known and observed costs. The savings from reduced crime rates, for example, can be measured in terms of reductions in expenditures on the criminal justice system. Such analyses are less scientifically sound than CBA since they do not attempt to value *all* relevant outcomes, in particular final outcomes for the individuals involved. One example of a cost-savings analysis is an evaluation of a London based Link Worker scheme. The scheme aims to improve access to support in the community for people with mental health problems who have come into contact with the criminal justice system (Revolving Doors Agency, 2000). This study looks at whether the costs to local agencies of providing additional services to the scheme's client group are offset by a fall in costs of 'crisis' services, including temporary accommodation and A&E services.

Principles of economic evaluation

Economic evaluation is underpinned by a number of fundamental principles derived from welfare economics. Welfare economics is concerned with the analysis of conditions under which policies may be said to have improved societal well being relative to alternative courses of action. The key words are: *relative*, *societal*, and *well being*. Evaluations should be comparative in nature, societal in scope and concerned with the resulting well being of the individuals involved. Not all evaluations will achieve all the aims of welfare economics, but this is the ideal against which evaluations should be judged and the standard for researchers to aim for.

Since economics is concerned with resource scarcity, economic analysis, by definition, must involve the comparative analysis of alternative courses of action (Drummond et al, 1997). In order to assess the true costs and benefits of an intervention, the most appropriate comparator should be the 'next best' alternative (Siegal et al, 1997), but where this is not clear-cut, a number of alternatives may be required, including perhaps the most widely practised alternative and current local practice, where these differ (Byford & Palmer, 1998). In some circumstances, the next best alternative may be 'do nothing', where no alternative or no proven alternative exists. If, for example, the only available comparator has not been assessed in a previous evaluation and is in fact cost-ineffective, then the study intervention will appear misleadingly cost-effective in comparison. Achieving a suitable comparison will depend on the nature of the intervention to be studied and the question to be answered. A comparison can only be undertaken, for example, if there is an appropriate group to compare an intervention with. Thus, universally imposed interventions, such as the new arrangements for all young people leaving local authority care or accommodation (Department of Health, 1999), will be more difficult to evaluate than a discrete

service set up as an addition to, rather than a replacement for, alternative services attempting to meet the same need.

The principles of welfare economics require an economic evaluation to adopt a societal perspective that includes all the possible costs and benefits of an intervention to all sectors in society (Johannesson, 1995). The perspective should be broad enough to include any indirect effects, such as the effect on carers, and long enough to capture all significant future effects. A full societal perspective can only be achieved where all significant impacts are known and where time and resources allow the gains and losses to be recorded. An intervention that has a limited number of relatively obvious impacts on a discrete and readily identifiable group of individuals or organisations will, therefore, be much simpler to evaluate from a societal perspective than an intervention that has significant, wide-ranging, long-term and perhaps hidden impacts.

The fundamental aim of welfare economics is to maximise societal well being, given existing resource constraints. Thus the focus of an economic evaluation will often be on final outcomes, such as quality of life, rather than intermediate outcomes (e.g. cholesterol level) or process measures (e.g. days in hospital or days in care), although these may be useful in helping to interpret changes in final outcomes. In addition, generic measures of quality of life enable comparisons across a much wider set of interventions than disease-specific or condition-specific outcome measures, as discussed above.

Nature of interventions

The ability to achieve the 'ideal' economic evaluation is influenced by the nature of the intervention under investigation. Interventions vary considerably along a spectrum of complexity and the more complex the intervention, the harder the evaluation will be and the harder it will become to follow the principles outlined above. This section examines characteristics of interventions that have particular implications for evaluation and which help to distinguish the more straightforward from the very complex, drawing out the difficulties of applying traditional economic evaluation techniques to relatively complex areas of health and social welfare.

Degree of user involvement

Interventions differ in the extent to which service users are 'active' or 'passive' in the delivery of the intervention – the extent to which they interact with the provider and how far the success of the intervention is dependent on their active response. The more actively involved service users are, the more likely it is that outcomes will be affected by individuals' values, culture, attitudes and circumstances. This introduces a set of confounding variables that need to be taken into account in an evaluation, but which are harder to categorise or quantify in the same way as, say, age or gender. A health promotion campaign would lie at one end of the spectrum since its effectiveness depends on whether people accept the message (e.g. smoking is bad for your health) and modify their lifestyle as a result (e.g. give up smoking). Patients undergoing emergency surgery are more passive. Although they can influence the recovery period, they can have little impact on the success or otherwise of the surgical procedure itself.

Many medical and surgical interventions require some patient involvement, including participation or compliance with the treatment. But, it is probably fair to say that this is not generally of the same order of magnitude as for more diverse community interventions. Receptiveness to a postnatal support service, for example, will depend on whether the intervention is seen as helpful or as an interference, how much support is received from friends and family, the quality of the relationship with the support worker, and so on – all of which will differ in subtle ways from one mother to the next (Morrell et al, 2000). A traditional quantitative economic evaluation comparing two groups in a randomised controlled design, would be able get a reasonable approximation of general cost-effectiveness, but may hide differential effects resulting from user-focused variables such as these.

A high degree of service user involvement and influence does not make economic evaluation in a quantitative framework impossible, but the level of heterogeneity does make for a more difficult evaluation. In order to explore and control for the influence of users, it is likely that a larger sample will be needed with the collection of a greater number of user-focused variables than would be the case in a more 'passive' group of users. Qualitative research can also be used to explore the influence of users and help to explain the results of the quantitative analysis (discussed in more detail below).

Complexity of intervention

Some healthcare interventions are highly standardised. A 200mg Ibuprofen tablet is just that. Some interventions, such as surgery, ought to be fairly uniform, but may vary significantly in practice, because of differences in clinical practice between consultants or hospitals and differences in the skills or competence of those delivering the service. Other interventions, such as community mental healthcare, are more complex or purposively flexible to match the varying needs of individual service users (Taylor & Thornicroft, 1996). Case managers, for example, whether health or social services based, provide a number of different services including advocacy, needs assessment, planning, implementation and co-ordination of services to meet needs. All these roles will be carried out in a variety of ways, to varying degrees, dependent on the needs of the clients, who may need support in many areas of life, including housing, employment, finances, medication or day care, to name but a few. Most social welfare interventions are similar in this respect.

Knowledge of the extent and nature of such variation is essential to understand how variations in the intervention affect outcomes. If, for example, an intervention appears to be more effective for some service users than others, it is helpful to know whether this is because some users were more receptive to the treatment or because the treatment itself was administered differently. Equally, when an intervention consists of a package of measures, it is useful to know whether some aspects of the package were having more of an impact than others.

Where variations can be measured in some way (e.g. drug dosage), it is possible to allow for these differences within a quantitative evaluation. However, there are limits to the number of different variables that can be incorporated. For example, postnatal support workers were required to keep a log of how much time they spent on each activity with each individual client (Morrell et al, 2000). In theory, this could be used to assess the effectiveness of each care component. In practice, however, the sample was too small to carry out a statistical analysis of

this kind. In other situations, variations that may be important to the success of the intervention may not be amenable to measurement or may be harder to observe – for example, the quality of the service provided by postnatal support workers, as assessed by the mothers.

Complexity of outcomes

For some healthcare interventions, outcomes are relatively easy to quantify. Many are judged by their success in reducing mortality or disease-specific morbidity that is relatively amenable to quantitative measurement, such as blood pressure level or symptom reduction. This is particularly true for conditions that are localised, physical and readily treatable, such as ingrowing toenails or broken bones. Other healthcare interventions seek to make a more 'holistic' impact on conditions that influence many aspects of a person's life, such as mental health problems. Such impacts are much harder to measure and compare with accuracy, since they are less amenable to objective measurement and can be multiple in nature, affecting psychological, social, family and physical functioning. The broader social welfare field is similar; aims and outcomes of social welfare interventions are often multiple and subjective in nature. Accurate representation of changes in outcome can, therefore, be difficult.

The solution should be simple – the ideal would be to focus on a measure of overall well being or quality of life, the focus of welfare economics. The problem, however, lies in the ability of current generic measures to capture the full range of impacts that interventions may have on an individual's life. In the healthcare field, much research has been carried out into the development of quantitative measures of outcomes, usually based on short questionnaires that enable a particular outcome to be scored on a standard scale. Provided they have been validated for the purpose for which they are being used, this allows all sorts of outcomes to be incorporated within a quantitative framework. But, this is an important qualification. Where appropriate validated measures do not exist, there is a danger that quantitative scales will not be measuring what evaluators believe they are measuring or that they will not be sensitive to the levels of change expected as a result of a particular intervention.

The majority of scales available that fulfil the needs of an economic evaluation (single index utility measures), such as the EuroQol (EuroQol Group, 1990), have been developed to measure health-related quality of life and tend to focus more on physical than psychological or social functioning. Although such scales have been successfully applied in the healthcare field, they may not be sensitive enough to measure change in broader social welfare contexts. The measurement of outcomes in more complex areas of mental health and social welfare can thus be constrained by the lack of appropriate scales.

Perspective and scope

The effects of many healthcare interventions are experienced by a limited and known number of organisations and groups in society and thus a societal perspective can be achieved with the inclusion of relatively few viewpoints. In comparing the treatment of ingrowing toenails by podiatrists or surgeons, for example, it would be reasonable to focus on the health service and the patient. A comparison of community and hospital-based care for elderly people with mental health problems, however, would need to be broader, perhaps including the health service, social services, voluntary services, informal care and costs borne by both the patient and the family.

Evaluation of social welfare interventions are particularly likely to span multiple sectors, since many different agencies are involved in providing services to vulnerable groups targeted by these programmes.

Time scale

Health and social problems vary enormously in terms of duration. Some relatively minor health conditions, such as a broken bone, are usually easy to fix. Similarly, some social care concerns can be relatively easily dealt with, such as the provision of meals for an elderly person or overnight accommodation for a child whose single mother needs to spend a night in hospital. Such 'acute' problems are short-term in nature and both the costs and effects of interventions are felt soon afterwards. Evaluation can thus be relatively short-term with little danger of missing important effects. More complex health and social problems, however, can be chronic and enduring; thus costs and benefits may need to be recorded over a much longer period of time, which increases the cost of an evaluation. An alternative solution is the use of shorter-term 'proxy' measures of outcome that suggest outcome movements in the right direction, such as young excluded people returning to school or families in difficulty attending family therapy sessions.

Summary

Social welfare interventions, like healthcare interventions, cover the whole spectrum from simple to complex. However, the majority have more in common with the treatment of chronic and enduring health problems than with the acute healthcare field. Social welfare problems are often long-term in nature and impact upon many areas of a person's life and the lives of their families. Interventions are characterised by a high degree of user involvement, significant variability across interventions and recipients, multiple and complex outcomes and multiple agency involvement. The evaluation of such interventions can be complex, hampering the ability to achieve the fundamental aims of welfare economics described above.

Although none of the complexities discussed preclude the achievement of a good quality economic evaluation, they do add significant difficulties and evaluations often fall short of the 'ideal' due in part to the difficulties themselves, but also due to the need for relatively greater funding that may not always be available. Evaluations should be designed with these characteristics in mind and efforts made to ensure that appropriate outcome measures are used, relevant sectors included, adequate follow-up periods are incorporated, and methods for controlling for and recording variation and compliance are developed.

Assessment criteria

As well as complying with the fundamental principles of welfare economics, an economic evaluation should be based on accurate and unbiased evidence of costs and effects. Such evidence can be obtained from a range of sources including randomised controlled trials, quasi-experiments, observational data, and qualitative research. Health economists are guided in their choice of method by a number of general criteria, although the interpretation of these criteria and

the importance attached to each will depend on the nature of the intervention and the purpose of the evaluation. In addition to feasibility, the two principle criteria are internal and external validity (Moser & Kalton, 1993).

Internal validity

Internal validity describes the extent to which the method of evaluation selected produces unbiased estimates of costs and benefits. Internal validity is achieved if observed differences in outcome or cost are due only to the difference in intervention received. The presence of confounding factors (other factors that may influence outcomes) reduces the internal validity of an evaluation study and efforts should be made to minimise such biases. In the healthcare field, internal validity is usually the over-riding criterion and should clearly be a major consideration in the evaluation of any other social welfare intervention, since an evaluation is of little use if it does not contribute to a better understanding of what additional impact an intervention has.

Some of the characteristics of complex health and social welfare interventions, discussed above, can hinder the achievement of internal validity. First, differences in outcome can be difficult to quantify due to their multidimensional nature. Secondly, social welfare interventions are beset by confounding factors, including the heterogeneity of the interventions themselves. In addition, social welfare interventions are likely to be part of a package of measures designed to reinforce one another, rendering care extremely variable and individual in nature. Statistical analysis is often used in the healthcare field to control for variation and the influence of confounding factors. The ability of statistical analysis to control for the scale of variability found in more complex areas of social welfare, however, may be limited. Furthermore, the need to observe and record these variations in order to help explain observed outcomes can render the concept of controlling for variation irrelevant. Thus, additional ways of satisfying the validity criterion (or the problem of "causal attribution") may need to be explored, perhaps involving greater reliance on qualitative research methods and theory-based approaches to evaluation (Sefton, 2000).

External validity

External validity, or generalisability, is concerned with how representative the sample is of the real world situation - how far the results are applicable to other user groups or settings. Whilst it is important for an evaluation to demonstrate how effective an intervention is in a particular setting, it is also important to assess how effective the same intervention would be in other settings – a different area, population or time period. To some extent, there is an inevitable trade-off between internal and external validity. An experiment that is carried out in tightly controlled conditions (e.g. a standardised service delivered to a homogeneous population) reduces the number of confounding factors, but also reduces the likelihood that the results will hold under more 'usual' circumstances.

In the healthcare field, concerns about external validity have tended to concentrate on artificial conditions and the representativeness of the sample. Some critics have argued that experimental trials are conducted under 'laboratory' conditions, so the results may not reflect what would happen in more 'natural' settings (Schwartz & Lellouch, 1976). In addition, patients and clinicians may behave differently knowing they are part of a trial – the so-called Hawthorne effect (Moser & Kalton, 1993). As an example, the authors of an economic evaluation of

misoprostol, an anti-ulcer drug, argued that compliance rates and detection of 'silent' ulcers were likely to be higher in the trial than in regular practice (Hillman and Bloom, 1989). For these reasons, the trial probably over-stated the effectiveness of the treatment and adjustments were made to generate estimates of cost-effectiveness that would more closely reflect regular practice. Increasingly, pragmatic or naturalistic trials are being carried out, where efforts are made to ensure that patients included in the trial are typical of normal caseload and that the treatment and follow-up of patients is conducted under routine conditions (Roland & Torgerson, 1998).

The same concerns are relevant to other areas of social welfare, where it is also important to ensure that the results of pilot programmes implemented in a single area, for example, are not automatically assumed to hold for other areas. The issues can be challenging because social policies are often highly context-dependent. If a particular drug treatment can be shown to work for a representative sample of adult men in the south of England, then it may be reasonable to assume that it will also be effective for adult men in the south, assuming similar rates of compliance with treatment. If, on the other hand, a specific regeneration initiative has been shown to work in one area, it is not safe to assume that it will automatically work elsewhere.The success of the scheme will depend upon the socio-demographic characteristics of residents and the underlying causes of problems, which will vary across areas.

The external validity criterion is thus a more difficult one to satisfy in more complex areas of social and healthcare. For many social welfare interventions, it may not be the results that are transferable, but knowledge about why certain interventions work under certain circumstances and for certain people (Pawson and Tilley, 1997). The focus is not so much on the estimation of the overall effects of a particular programme, but on identifying successful mechanisms within the programme and the conditions necessary to trigger them. This in turn has implications for the best way to go about an evaluation.

Relevance

As well as being valid, an evaluation must be relevant, i.e. provide evidence that is useful to decision-makers in pursuit of their aims. Conventional approaches to economic evaluation in the healthcare field are designed to answer one particular question: 'Is the intervention an efficient use of resources compared with alternative uses?' and the guidelines for economic evaluation tend to be centred around this one, albeit important, question. This approach is most suited to the evaluation of established interventions where the only question decision-makers are concerned with is whether the intervention is effective and, if so, whether it is cost-effective compared to existing treatments.

Some healthcare and many social welfare interventions evolve over time or are purposely flexible to meet the changing needs of service users. Therefore, other evaluation questions may be more relevant. In particular, evaluation has an important role to play in the learning process, by looking at how a scheme might be improved or developed over time. For an economist, this means identifying ways in which a scheme could be made more efficient, requiring a more dynamic approach to evaluation. In the healthcare field, particularly more complex areas such as community mental healthcare, evaluators have tended to rely upon additional evaluations as the care provided develops. But, there is an important difference between an evaluation that assesses

the cost-effectiveness of a modified programme and one that helps to inform what changes should be made to the programme to increase its cost-effectiveness.

Study design

The appropriate choice of study design is guided by what is feasible and what is most likely to produce results that are internally valid and generalisable. This section looks at some of the preferred research designs used in the healthcare field and their potential applicability to other areas of social welfare. In practice, an economic evaluation may combine several methods within the same study.

Randomised controlled trials

In the healthcare field, evaluations are increasingly carried out within randomised controlled trials (RCT) and emphasis placed on quantitative data (Pocock, 1983). Participants are randomly assigned to either an experimental intervention or the control intervention. Ideally, costs and outcomes for both groups are monitored prospectively over time and compared. Differences between the two groups are attributed to the intervention since the randomisation process is designed to ensure that estimates of effects are unbiased by the selection of participants, thus internal validity tends to be high. For this reason, evidence from RCTs is rated more highly than evidence from other sources, to the extent that systematic reviews often exclude other designs (NHS CRD, 1996). Other methods tend to be judged by how closely they approximate the 'gold-standard' RCT.

RCTs are relatively more prevalent in the healthcare sector than in other sectors. This is mainly a desire for scientific rigour, but requirements from policy-makers, journal editors and research funding bodies have also played a part. Some of the healthcare programmes that have been evaluated using RCTs have much in common with the wider social welfare field. Examples include the evaluation of postnatal support for new mothers (Morrell et al, 2000), education programmes for people with osteoporosis of the knee (Lord et al, 1999), diet-exercise programmes for people with diabetes (Kaplan et al, 1988) and community care packages for people suffering from a mental illness (UK700 Group: Byford et al, 2000).

These interventions raise many of the issues discussed in previous sections. They are not standardised services delivered to 'passive' participants and many of the outcomes are not easily quantified. Whilst this shows that it is feasible to carry out RCTs for more complex interventions, this does not necessarily mean that it is the best approach to adopt, at least not without addressing some of the problems raised. For illustrative purposes, some specific issues are discussed in relation to the evaluation of a postnatal support service for new mothers (Morrell et al, 2000). This study, summarised in Box 1, is a scientifically sound and rigorous experimental evaluation, yet the results raise many questions about the limitations of the experimental approach in the context of a relatively complex and heterogeneous health service.

Box 1: Economic Evaluation of a Postnatal Support Scheme.
Study design: Randomised, controlled trial comparing usual community midwife care

with usual care plus access to a community support worker for up to ten visits within the first 28 days after the birth.
Sample: 623 new mothers aged 17 years or over.
Main outcome measures: Primary measure – the general health perception domain of the Short Form-36. Secondary measure – Edinburgh Postnatal Depression Scale, Duke Functional Social Support Scale and breastfeeding rates.
Method of economic evaluation: Cost-consequences analysis.
Results: No significant differences were found in any of the outcome measures although there was a trend suggesting that the control group performed better on some of the secondary outcome measures. Satisfaction with services was higher for those in the experimental group. The total mean cost to the NHS of the experimental group was estimated to be £180 per woman greater than for the control group. No differences were found in the use of NHS services, excluding the cost of the intervention.

- *Reliability of quantitative outcome scales* – Like other similar studies, this evaluation relied upon quantitative scales to measure such things as the mothers' general health and levels of depression. The validity of the results are only as good as the validity of the scales, so any claims need to be tempered by a note of caution about the appropriateness of these outcome measures for the purpose of this study. In fact, health outcomes after six weeks showed no positive effects and some negative effects for those who received the support service as compared to those who did not, yet participants were overwhelmingly positive about the scheme. The outcome measures may be accurate, but it is possible that the measures used were not picking up those outcomes of importance to mothers. Although this ambiguity might be resolved by further studies using different outcome measures, it demonstrates the possible limitations of using quantitative scales, even where these have been validated.

- *Focus on net outcomes* – Like other RCT-based studies, this evaluation focused on net outcomes, without providing much information on how, for whom, and in what circumstances the intervention worked. In some situations, it is important to understand how a programme is working because this helps to assess whether the same intervention would be effective in other settings and how the scheme might be improved in the future. In this evaluation, the authors suggest that the service might have failed because it did not build up people's own coping skills. This might explain why outcomes after six weeks were not positive, but, as the authors admit, this theory is speculative. The evaluation was unable to provide any evidence as to why the scheme did not appear to be effective, yet this is crucial to the overall assessment of the scheme.

- *Timing of outcomes* – Like other studies, outcomes were measured at discrete points in time – after six weeks and six months. The timing of surveys is often arbitrary, although this can have a significant impact on the results. In this case, the first survey was undertaken shortly after the intervention ceased and the authors later recognised that mothers may have been experiencing a 'withdrawal' effect, which could explain the negative outcomes for the intervention group. However, it is not possible to know this for sure or to know how long the withdrawal effect lasted, if indeed, that is what it was.

- *Standardisation* – Even though the service consisted of a number of different activities and although the amount of time spent on each activity varied from one mother to the next, an RCT design effectively treats the intervention as if it were a single homogenous service. It is debatable whether a clinical trial that prescribed patients with a variable cocktail of drugs would be considered valid, and this is really no different. Even if it were valid, this approach is unable to address questions about the appropriate mix or timing of activities without a number of sub-group analyses, requiring a larger sample size.

As it stands, the evaluation suggests that the support scheme was not cost-effective. Carried out differently, however, another evaluation might have reached a different conclusion. Had outcomes important to mothers been determined, perhaps using qualitative assessment before or during the trial, 'hidden' benefits of the intervention that would explain the positive reaction to the scheme may have been discovered. In addition, a more flexible design with a longer-term focus that allowed the scheme to evolve (e.g. by allowing the support to be extended or phased out gradually) may have been of benefit.

These issues are not criticisms of RCTs per se, but of the relatively rigid nature of scientific trials as commonly designed in the healthcare field. Traditional RCT designs do not provide much information on how or why an intervention is working, nor are they particularly helpful in identifying ways to improve or build upon the success of a particular scheme. They are more suited for carrying out a summative evaluation of defined and established interventions. These weaknesses do not preclude the use of (more flexible) experimental designs, since an unbiased sample should always be an important goal. There is no doubt, however, that the cost of large RCTs would preclude the possibility of testing an intervention every time it evolved in some significant way, thus, other designs are likely to be more important in earlier stages of development.

Evaluators should be aware of the potential pitfalls of following a 'traditional' RCT model and be open to the additional benefits that can be gained from pragmatism and extensions to the model such as qualitative analysis, pre-trial testing and pilot studies. Social welfare programmes, as discussed in section 2, tend to be complex, supporting a wide range of people with a spread of individual needs and circumstances that may influence outcomes. In order for an RCT type design to be usefully employed, such variations must be explored in sub-group analyses or controlled for in multivariate regression analyses, often requiring a greater data collection burden and larger sample sizes than would be the case where variation is more limited.

There are, in addition, some genuine obstacles to RCTs in the social welfare field that may on occasion preclude such designs, and a greater understanding of these obstacles may help to overcome them:

- *Ethical concerns* – Since randomisation involves denying treatment to some participants, concerns are often expressed about the ethics of randomisation, a debate that will be familiar to many health economists (e.g. Hafner & an der Heiden, 1991; Williams, 1992). In defence, there is nothing unethical about restricting services to one group of people, where the impact of that service is as yet unproven, and indeed may be harmful. RCTs thus tend to be used to evaluate new interventions that people would not necessarily have access to in the absence of

an evaluation. However, there are situations where ethical concerns may result in a decision not to randomise. For example, where treatment is for a life-threatening condition or where services require considerable motivation, such as self-help initiatives (Brewin & Bradley, 1989; Bagenal et al, 1990; Sheard, 1990). In the social welfare field, ethical objections to randomisation will often be for reasons of safety. One example is local authority care and accommodation where it would undoubtedly be unethical to randomly allocate a young person to return home to their birth parents if that young person had been removed from the home due to serious neglect.

- *Practical problems* – Evaluators may not be in a position to decide who does and who does not receive a particular treatment. Service providers may put other priorities ahead of evaluation; for example, they may want to target their services at potential clients who they think will benefit most, rather than use random assignment. On occasion, national policies are implemented untested and country-wide, such as the new arrangements for young people leaving care (Department of Health, 1999), making it impossible to construct a comparison group who are not affected by the intervention.

- *Cost* - A well-conducted RCT can be expensive and the necessary resources required may not be available. Social welfare research, being more focussed on qualitative sampling and without the push for sample sizes capable of producing statistical significance, may be less expensive than scientific RCTs. Thus, to encourage RCT designs requires the understanding of research funding bodies perhaps more familiar with less resource intensive designs, as well as the understanding of the evaluators themselves. Of course, relatively small and cheap RCTs can be carried out, but they may only be able to pick up relatively large differences in outcomes and the scope for sub-group analysis will be more limited.

Non-randomised experiments

Non-randomised experiments are an attempt to approximate RCTs in situations where randomisation is not feasible or is inappropriate for some of the reasons already discussed. They share many of the same characteristics as RCTs, but with some important differences. Because service users are not randomly assigned to different treatment groups, there is a danger that the results will be biased, threatening the validity of an evaluation. The potential for bias depends on how participants are allocated to different interventions. In non-randomised experiments, participants may be offered the intervention that is considered to be most appropriate to their individual needs. In this case, outcomes would be expected to be greater than for a random sample of the target population. Alternatively, those perceived to be most in need may be allocated to the 'experimental' intervention and thus their outcomes may be expected to be worse. In some cases, service users volunteer or self-refer for particular interventions. Again, we might expect those who stand to benefit most from an intervention and/or those who are more enthusiastic or motivated to be more likely to volunteer.

Some non-randomised experiments involve a straight comparison between an intervention group and a comparison group. Quasi-experiments, on the other hand, seek to minimise any potential bias by controlling for known differences between groups. The most straightforward approach is to 'match' the groups in terms of their socio-demographic and other characteristics that are

expected to influence outcomes. A more sophisticated approach is to use statistical analysis to adjust for differences in the characteristics of the groups. Whilst some adjustment is better than none, quasi-experiments can only control for observable differences between the intervention and comparison groups. Some important differences may not be controlled for, if they are difficult to identify or measure and, in any case, there may be a limit to the number of characteristics that can be controlled for because of sample size.

An evaluation of on-site physiotherapy illustrates the potential problems inherent in non-randomised experiments (Hackett et al, 1993). Costs and outcomes were compared across three rural practices, each of which operated a different referral system (on-site physiotherapy, access to physiotherapy at a local hospital or referral to orthopaedic consultants). Patients could not be randomly assigned because the referral system was determined by the practice attended, but the practices were chosen because the demographic profile of their patients and the prevalence rates for joint and soft tissue injuries were shown to be similar. Thus, it was effectively a quasi-experiment, based on matched controls. The study showed that at the on-site physiotherapy practice, patients were treated more quickly, spent less time off work, had fewer financial costs, and fewer prescriptions. However, referral rates were much higher at this practice. One possible explanation is that patients were being treated for more minor injuries. The authors offered an alternative explanation, but there is still at least potential for bias in the results.

In the healthcare field, non-randomised experiments are generally seen by economists as a second-best alternative to RCTs because the results are more likely to be biased. What is often not discussed is whether second best is good enough. In other words, are the results of a non-randomised experiment likely to be sufficiently valid to base decisions about resource allocation upon? This depends on the scope for bias, how far this can be controlled and whether the resulting bias is large relative to the expected effects of the intervention being evaluated.

In the social welfare field, non-randomised experiments are often considered to be a good compromise without proper consideration of the issues. In the case of welfare-to-work, some commentators have argued that the results of quasi-experiments are generally unreliable and can be misleading (e.g. Lalonde, 1986). Some of the potential sources of bias are unobservable and, therefore, cannot be controlled for (e.g. the motivation of participants). On top of this, these schemes tend to have relatively low-level effects, so small amounts of bias can affect the results significantly. Where evaluations are based on a straight comparison between an intervention group and a control group, evaluators need to be more aware of potential bias and, where possible, efforts should be made to reduce this bias, by careful matching. Economic evaluations in the healthcare field often include a discussion of possible sources of bias and the same practice could usefully be followed in other areas of social welfare.

Modelling

Modelling can be a useful tool for evaluators. All models involve a set of inputs and assumptions and a more or less complex process for generating outputs. At its simplest, a model may be used to test the sensitivity of the results to changes in key parameters (e.g. medication compliance rates). More complex decision-analysis models can involve simulating medical histories for synthetic cohorts of patients and estimating intervention costs and outcomes under various treatment strategies (Sculpher et al, 2000). Models are used in a range of situations:

- As a scoping exercise to help determine which variables are likely to be of greatest importance and should be given most attention in designing an RCT.

- To generalise the results of an evaluation, either from trials to regular practice (e.g. adjustment for non-compliance) or from place to place (e.g. adjustment for differences in referral rates or unit costs).

- To extrapolate beyond data on intermediate outcomes in order to estimate final outcomes (e.g. life expectancy based on reduced short-term mortality or morbidity) or to extrapolate beyond the time period covered by an evaluation.

- To combine data from a range of different sources within a uniform framework. For example, the Stroke Prevention Policy Model combined RCT data on short-term clinical outcomes with observational data used to extrapolate life-expectancy, preference surveys to adjust for quality-of-life, and administrative data on the use of services by patients in different conditions (Matchar et al, 1997).

Modelling is an important element of many economic evaluations in healthcare and is often used to complement other methodologies. It has the advantage of flexibility and is particularly useful as a way of overcoming some of the limitations of RCTs, helping to extrapolate longer-term outcomes and generalise the results to other settings. But, the outputs of a model are only as good as the quality of the inputs, so models are not a substitute for a lack of data, although they can help to make the best use of the data available (Sculpher et al, 2000). In many areas of social welfare, the scope for modelling will be constrained by lack of data. In these situations, a different approach is needed.

Qualitative methods

Qualitative research covers a wide range of techniques, including in-depth interviews, focus and consensus groups, and case studies. Unlike the methods discussed above, the aim is not to provide quantified answers, but to focus on the meanings, experiences, views and motivations of participants. As such, qualitative research is useful for addressing a different set of evaluation questions: not how many “X”s, but what is “X”, how does “X” vary in different situations and why? As one commentator put it, “Healthcare [and other social welfare services] deals with people and people are, on the whole, more complex than the subjects of natural science. There are a whole set of questions about human interaction and how people interpret interaction which experimental and other quantitative methods are less suited to answer.” (Pope and Mays, 1995)

There are a number of ways in which qualitative methods can be used to complement quantitative approaches to evaluation:

- As a preliminary step to help determine the scope and design of a quantitative study (e.g. identifying outcomes of interest or the best way of measuring them).

- Helping to interpret the results of quantitative studies (e.g. why certain outcomes are observed) or to assess the process of evaluation (i.e. whether a programme is being delivered in the way it is intended to be).

- As a substitute for quantitative evaluation in areas that are less amenable to quantitative research (e.g. organisational change).

At present, qualitative research methods are not widely used by economists in the healthcare field. One of the reasons, other than lack of familiarity, is a concern about the scientific rigour of these methods. This results in part from a misunderstanding about the purpose of qualitative research. The aim is not to generate statistically significant results, but to explore themes, patterns, and associations within a richer and more diverse data set. Where a quantitative researcher would attempt to show causality by looking for statistical correlations between variables, a qualitative researcher would look for explanations within a qualitative dataset.

Used appropriately, qualitative research can help to address some of the shortcomings of RCTs and other quantitative approaches highlighted in our critique of the postnatal support service. For example, qualitative research could have been used to help identify those outcomes important to mothers and to ensure their validity and comprehensiveness. It could also have been used to understand how the service affected mothers beyond the intervention period. For example, was it the case that supported mothers felt less able to cope when the support was withdrawn? Finally, it could have been used to explore ways of improving the effectiveness by assessing whether some elements of the service were valued more highly than others or exploring why some mothers declined the service.

Concluding remarks

Throughout this paper a spectrum of complexity in interventions has been implied. Although we do not argue that healthcare interventions are ‘simple’ whilst other social welfare interventions are ‘complex’, it is reasonable to suggest that many social welfare interventions demonstrate a level of complexity that is reached by few healthcare services.

The problems that may be encountered when applying ‘traditional’ economic evaluation techniques, highlighted in this paper, are by no means exclusive to areas of social welfare other than healthcare. Indeed, within healthcare, rigid reliance on randomised designs, quantitative data and statistical validity is noticeably more common in relatively standardised and homogeneous areas of healthcare, such as surgery and pharmaceuticals, demonstrating the relative ease of scientific evaluation in some areas as compared to others. But the difficulties are likely to be more extensive in the broader field of social welfare, so perhaps it is no surprise that alternative methodologies are more heavily relied upon in these areas.

This is no reason for techniques preferred by health economists to be dismissed. There are very strong arguments for randomisation and demonstration of statistical significance in evaluations, which are not exclusive to the healthcare field. All evaluations should aim to provide answers in

which we can have confidence and the likelihood of this will be increased through attention to unbiased sampling, appropriate outcome measures, generalisability and statistical significance.

The lesson to be drawn from this paper is that efforts need to be made to ensure greater scientific validity of evaluations, whilst recognising that the complexity of many social welfare interventions may require greater attention to the details of trial design and modifications or additions to the 'conventional' model. Although movements away from the 'ideal' will undoubtedly have a negative impact upon the internal validity of results, this must be balanced against the likelihood of meaningless results from the imposition of rigid scientific techniques that are inappropriate, given the nature of the interventions under evaluation.

In general, the more complex an intervention, the more likely it is that an experimental approach to economic evaluation will not be sufficient on its own. There is a greater role for qualitative research methods, in particular, in helping to address many of the limitations of experimental approaches; for example to check the appropriateness of outcome measures or to corroborate the findings from the quantitative analysis. The results of economic evaluations also need to be presented and interpreted more carefully than is often the case, acknowledging weaknesses in the study design and exploring potential biases that may result. Furthermore, greater attention to how and why programmes work can help strengthen the validity of a study, by specifying the links between the intervention and observed outcomes, and improve generalisability through a better understanding of the context in which a particular intervention is likely to be effective.

References

Bagenal FS, Easton DF, Harris E, et al (1990). Survival of patients with breast cancer attending Bristol Cancer Help Centre. *Lancet*, 336: 606-610.

Bala MV, Mauskopf JA, Wood LL (1999). Willingness to pay as a measure of health benefits. *PharmacoEconomics*, 15 (1): 9-18.

Brewin C, Bradley C (1989). Patient preferences and randomised controlled trials. *British Medical Journal*, 299: 313-315.

Byford S, Palmer S (1998). Common errors and controversies in pharmacoeconomic analyses. *PharmacoEconomics*, 13: 659-666.

Department of Health (1999). *Me, survive, out there? New arrangements for young people living in and leaving care*. London: Department of Health.

Drummond MF, O'Brien B, Stoddart GL, Torrance GW (1997*). Methods for the economic evaluation of health care programmes*. Oxford: Oxford Medical Publications.

Drummond M (1991) Common mistakes in the design of economic evaluations of medicines. *British Journal of Medical Economics*, 1: 5-14

EuroQol Group (1990). EuroQol - a new facility for the measurement of health-related quality of life. *Health Policy*, 16: 199-208.

Gafni A (1998). Willingness to pay: what's in a name? *PharmacoEconomics*, 14(5): 465-470.

Hackett GI, Bundred P, Hutton JL, Obrien J, Stanley IM (1993). Management of joint and soft tissue injuries in three general practices: value of on-site physiotherapy. *British Journal of General Practice*, 43: 61-64.

Hafner H, an der Heiden W (1991). Evaluating effectiveness and cost of community care for schizophrenic patients. *Schizophrenia Bulletin*, 17(3): 441-451.

Hillman A, Bloom B (1989). Economic effects of prophylactic use of misoprostol to prevent gastric ulcer in patients with arthritis taking non-steroidal anti-inflammatory drugs. *Archives of Internal Medicine*, 149: 2061-2065.

Johannesson M (1995). A note on the depreciation of the societal perspective in economic evaluation of health care. *Health Policy*, 33: 59-66.

Kaplan R, Atkins C, Wilson D (1988). The cost utility of diet and exercise interventions in non-insulin dependent diabetes mellitus. *Health Promotion*, 2(4): 331-340.

Lalonde R (1986). Evaluating the econometric evaluations of training programmes with experimental data. *American Economic Review*, 76: 604-620.

Loomes G, McKenzie L (1989). The use of QALYs in health care decision making. *Social Science and Medicine*, 28(4): 299-308

Lord J, Victor C, Littlejohns P, Ross FM, Axford JS (1999). Economic evaluation of a primary care-based education programme for patients with osteoarthritis of the knee. *Health Technology Assessment*, 3(23).

Matchar DB, Samsa GP, Matthews JR, Ancukiewicz M, Parmigiani G, Hasselblad V, Wolf PA, D'Agostino RB, Lipscomb J (1997). The Stroke Prevention Policy Model: linking evidence and clinical decisions. *Annals of Internal Medicine*, 127(8): 704-711.

Mauskopf JA, Paul JE, Grant DM, Stergachis A (1998). The role of cost-consequences analysis in healthcare decision-making. *PharmacoEconomics*, 13(3): 277-288.

McDaid D, Sefton T, Byford (2002). Economic evaluations in social welfare: a review of the literature. LSE Health and Social Care Discussion Paper (forthcoming).

Morrell CJ, Spiby H, Stewart P, et al (2000). Costs and benefits of community postnatal support workers: a randomised controlled trial. *Health Technology Assessment*, 4(6).

Moser CA, Kalton G (1993). *Survey methods in social investigation*. Aldershot: Dartmouth Publishing Company.

NHS CRD (1996). *Undertaking systematic reviews of research on effectiveness: CRD guidelines for those carrying out or commissioning reviews*. NHS CRD Report No 4.

O'Cathain A, Brazier JE, Milner PC, Fall M (1992). Cost-effectiveness of minor surgery in general practice: a prospective comparison with hospital practice. *British Journal of General Practice*, 42: 13-17.

Pawson R, Tilley N (1997). *Realistic Evaluation*. London: Sage.

Pocock SJ (1983). Clinical trials: a practical approach. New York: Wiley.

Pope C, Mays N (1995). Reaching parts other methods cannot reach: an introduction to qualitative methods in health and health services research. *British Medical Journal*, 311: 42-45.

Revolving Doors Agency (2000). *Mental Health, Multiple Needs, and the Police*. London: Revolving Doors Agency.

Robinson R (1993). Cost-benefit analysis. *British Medical Journal*, 307: 924-926.

Roland M, Torgerson DJ (1998). Understanding controlled trials: what are pragmatic trials? *British Medical Journal*, 316: 285.

Schwartz D, Lellouch J (1976). Explanatory and pragmatic attitudes in therapeutic trials. *Journal of Chronic Diseases*, 20: 637-648.

Sculpher M, Fenwick E, Claxton K (2000). Assessing quality in decision analytic cost-effectiveness models: a suggested framework and example of application. *PharmacoEconomics*, 17(5): 461-477.

Sefton, T (2000), *Getting Less for More: Economic Evaluation in the Social Welfare Field*. CASE paper 44. London: Centre for the Analysis of Social Exclusion.

Sen A (1985). *Commodies and Capabilities*. Amsertdam: Elsevier Publishers.

Sheard TA (1990). Bristol Cancer Help Centre (letter to the editor). *Lancet*, 336: 683.

Siegel JE, Torrance GW, Russell LB, et al (1997). Guidelines for pharmacoeconomic studies. Recommendations from the panel on cost effectiveness in health and medicine. *PharmacoEconomics*, 11: 159-168.

Sugden R, Williams A (1990). The principles of practical cost-benefit analysis. Oxford: Oxford University Press.

Taylor R, Thornicroft G (1996). Uses and limits of randomised controlled trials in mental health service research. In: Thornicroft G & Tansella M (eds), *Mental Health Outcome Measures*, Berlin: Springer-Verlag.

UK700 Group (Byford S, Fiander M, Torgerson DJ, Barber JA, Thompson SG, Burns T, van Horne E, Gilvarry C, Creed F) (2000). Cost-effectiveness of intensive v. standard case management for severe psychotic illness: UK700 case management trial. *British Journal of Psychiatry*, 176: 537-543.

Williams A (1992). Cost-effectiveness analysis: is it ethical? *Journal of Medical Ethics*. 18: 7-11.

Williams A, Kind P (1992). The present state of play about QALYs. In: Hopkins A (ed). *Measures of quality of life*. London: Royal College of Physicians.