Complexity, context and control in evaluating public health interventions: challenges for economic evaluation

Mark Pennington[1], Joanne Gray[2], Heather Dickinson[1], Cam Donaldson[1], Jan Walker[1]

1. Institute of Health and Society, Newcastle University
2. School of Health and Social Care, Teesside University

Paper presented to Health Economists' Study Group,
Brunel University, 5-7th September 2007

Name and address for correspondence:

Mark Pennington
Institute of Health and Society
Newcastle University
21 Claremont Place
Newcastle upon Tyne
NE2 4AA

mark.pennington@ncl.ac.uk

The evaluation which forms the basis of this paper has been funded by the Department for Children, Schools and Families, and we gratefully acknowledge their permission to discuss our experience of the evaluation.

1. Introduction

Economics is well established in the evaluation of health and health care. However, the boundaries of health and social care are becoming blurred as government health policies target lifestyle and environment in an effort to improve overall health and halt widening inequalities (DoH 2004). Perhaps as a consequence, health and wellbeing are increasingly viewed as multi-faceted concepts in which health care along with behavioural and environmental factors all have an important role to play. In parallel, the use of evidence in forming policy is now well established in the public sector; note the recent formation of the Campbell Collaboration as a sibling to the Cochrane Collaboration. This has highlighted the importance of evaluation techniques of effectiveness and efficiency, reinforced by the National Institute for Health and Clinical Excellence taking on a public health brief. Economists need to adapt evaluation techniques to address multi-agency settings with an increasing emphasis on social functioning.

The trend by public sector organisations towards commissioning based on local needs that are often context specific and thus dependent on the immediate social environments means that service provision across geographical areas is increasingly heterogeneous. Interventions in public health and social welfare often require the coordination of a number of actors, with manifestation of effects over long time spans. Their interaction with an evolving, heterogeneous environment challenges attempts to deconstruct these complex interventions when attempting to evaluate them.

Some of these challenges have been recognised in the literature on public health interventions (Kelly 2005). A number of authors have addressed the difficulties encountered in quantitative study design (Susser 1995; Campbell, Fitzpatrick et al. 2000; Wolff 2000; Hawe, Shiell et al. 2004; Shemilt, Harvey et al. 2004). However, challenges still exist in both the implementation of study methodology and the interpretation of results.

In this paper we address the above issues within the context of a specific and continuing study, funded by the Department for Children, Schools and Families (DCSF) (formerly Department for Education and Skills - DfES), evaluating the Budget Holding Lead Professional (BHLP) initiative for children with additional needs. The purpose of presenting to HESG colleagues is for us to learn from previous experiences elsewhere as to whether our arguments are valid and how to move forward from here, not only for this specific evaluation but also for future studies.

In the following section, we attempt a summary definition of complex interventions and provide more details on the BHLP initiative. We then attempt, in sections 3-5, to discuss three general areas of importance to evaluating such initiatives; how such interventions are interpreted and implemented by service providers; general issues of statistical design; and issues more specific to economic evaluations. In each section, we outline what might be ideally required from a research perspective, even if taking a reasonably pragmatic view, before outlining our experience to date in the

BHLP evaluation. Before offering some concluding comments, in section 6 we set out our proposed model for the quantitative evaluation, which we hope will help set an agenda for discussion.

2. BHLP initiative as a complex intervention

Complex interventions have been defined as those interventions that consist of a number of interconnecting elements that seem essential to the proper functioning of the intervention, but the mechanism through which this is achieved is uncertain (MRC 2000). As such, complex interventions are often programmes whose effects are crucially dependent on context and implementation. Indeed, Wolff characterises complex social interventions by their complex and diverse staffing arrangements, ambiguous protocols, unclear inclusion criteria, unevenly motivated subjects and dependence on the broader environment (Wolff 2000). A combination of high degrees of user involvement, individually tailored packages of care, heterogeneous recipients, multiple goals and inputs pose significant challenges for evaluators.

The BHLP initiative is typical of the complex social care interventions economists are called upon to evaluate. The initiative is part of a programme of changes in the delivery of targeted services to children with 'additional needs'. These are children who would not normally reach the threshold for statutory intervention by social services, but who are failing or likely to fail to meet one or more of the five key outcomes identified in the Every Child Matters agenda for change in children's services (DfES 2004). These outcomes are: being healthy; staying safe; enjoying and achieving; making a positive contribution and economic wellbeing. These children and their families are likely to be assessed as needing interventions from two or more agencies. A number of changes have been put in place to improve delivery of such packages of interventions including information sharing protocols (DfES 2006a), standardised needs assessment (DfES 2006b) and the appointment of a single agency worker (Lead Professional - LP) to coordinate assessment and service provision (DfES 2006c).

The appointment of a LP is part of a process of streamlining and coordinating the services families receive to reduce duplication and inconsistent support. The LP coordinates a team of appropriate agency staff to work with the child or family to address their problems. This process has been taken a step further with the introduction of Budget Holding Lead Professionals (BHLPs) in sixteen local authorities (LAs) in England (DfES 2006d). The pilot scheme, introduced by the DCSF (formerly DfES) in 2006, would give LPs access to and responsibility for a budget to procure services for the child or family. In principle, the policy represents a fundamental shift of power over service provision from agencies to their front line workers, and the children and families in their care.

The LAs were chosen by tender, and each was granted £525,000 over two years to implement the necessary support structures for BHLP provision and to pump prime the BHLP budgets. Long term funding support for BHLP is

intended to be obtained from core service budgets. Early intervention for children with additional needs is expected to generate long term savings by preventing deterioration in the child's situation and reducing demands on expensive statutory service provision.

It was evident from the beginning that the needs of these children, and the packages of care assembled, would be highly heterogeneous. The overall effectiveness of BHLP working is likely to be more than the sum of the intervention parts. A number of other factors may contribute to its effectiveness, not least the facilitation of communication between agencies and the child or family, leading to more appropriate and timely service provision and improved participation from the recipients. Effectiveness will also be governed by the training and preparation of staff and the service delivery structure, particularly the history of cooperation between agencies. The analysis framework for BHLP working shares many of the challenges of analysing complex interventions in that it is not clear which elements of the initiative are key to effective delivery.

The DCFS commissioned an evaluation of BHLP to be conducted alongside the pilot and report in March 2008. In addition to qualitative evidence of the impact of BHLP on service users and providers the evaluation would require quantitative evidence of effectiveness and cost-effectiveness of BHLP compared to LP working. The work was divided into two stages: first, an initial 'scoping study' (from September 2006 to January 2007), in which each site was visited by the study team and interviews conducted with key professionals; and the more in-depth qualitative and quantitative analyses to take place from January 2007 to March 2008.The original intention of the evaluation was a three arm comparison of traditional service provision, LP working and BHLP working, but it quickly became apparent that identification of children receiving services via traditional service provision in the sixteen pilot LAs was not possible as all were moving to the LP model as the standard way of working. Consequently the evaluation would compare LP and BHLP working.

3. Evaluation culture and 'evaluability'

Appropriate implementation of an intervention is essential if a meaningful evaluation is to be attempted. In health care research, staff delivering interventions under evaluation generally have an understanding of research methods, work in an environment where there is an established evidence based practice culture and, often, are members of research teams. The contrast with social care is stark. Evaluation methodology is less well established and often meets outright misunderstanding and distrust from professionals implementing pilot programmes (Oakley, Strange et al. 2003). In addition, the nature of complex programmes may impede control by researchers over key elements of the evaluation. Shemilt et al. highlighted that maintaining the integrity of evaluations of complex interventions is compromised by researchers having less control than is usual over the

process of recruitment, eligibility checking and the implementation of the relevant intervention (Shemilt, Harvey et al. 2004).

Our experience of evaluating the BHLP initiative has highlighted an unfamiliarity with quantitative evaluation research culture at local and possibly national level. LAs have engaged in a large number of evaluations in previous years, but these have been limited to qualitative evaluations and service user questionnaires. Policy elucidation was not clear enough and implementation guidance was insufficient. The original intention of the initiative was to devolve control over resources, and hence access to services, to the frontline professional. It was anticipated that this would facilitate a more comprehensive and responsive package of care for the child or young person. Control over a budget would encourage professionals to 'think outside the box' when confronted with diverse needs. The BHLP approach was implemented in most pilots, however, as a 'top-up fund' – a grant to pay for goods and services not currently available 'for free.

Whilst moves towards devolving control over resources to frontline professionals had been timid, the call to encourage professionals to 'think outside the box' was taken up with gusto. Professionals were using the fund to provide a number of goods for families which included landscape gardening and services such as beauty-makeover days to improve self esteem. Thus, it seems that BHLP money was simply turned into another budget line to dip into when required to purchase goods that could not be purchased through previously existing routes. ' The difference between policy intent and practice was unearthed during the scoping phase of the evaluation, and, at the point, officials and the Minister responsible gave clear signals to the pilots that they needed to move nearer to policy intent. Nevertheless, some six months later it was clear to the evaluators that little had changed. Officials again made concerted efforts to move the pilots closer to policy intent and we have yet to see whether this will be possible. Many pilots are reluctant to change direction given that they believe that the 'top up fund' is proving to be 'successful'.

Perhaps the most challenging cultural gulf relates to the purpose of the evaluation. The LAs approached the pilot as a 'teething stage' to iron out problems before rolling the approach out. In many LAs the pilot gathered momentum and they were encouraged to broaden the policy roll-out by external consultants. Consequently, after a year many BHLP managers did not appreciate our need to halt the roll out to facilitate the evaluation. Although the research tender had requested a comparison of LP and BHLP working, attempts to preserve some LP working in the LAs to facilitate this were thwarted: some LAs already regarded all LPs as BHLPs, even though most had not received any specific training for the new role; others were keen to allow both LPs and BHLPs 'access' to the top-up fund; and others had not implemented a LP approach prior to introducing BHLP working. Attempts to maintain a comparison group have been regarded by some LAs as the evaluators redefining policy intent or, at least, calling the implementation tune.

In order to overcome the confusion of stakeholders in implementing government policy, it has been argued that a more structured approach

should take place during the pre-evaluation phase (Kelly 2005). It has been suggested that preparatory phases are important not only in generating hypotheses, but also identifying likely sources of variability and confounding factors (Byford and Sefton 2003). Researchers, in their evaluation of complex community based initiatives surrounding health promotion, have advocated the application of theory based approaches to pre-evaluation (Judge and Bauld 2001); (Lafferty and Mahoney 2003). The philosophy behind this is that prior to the start of a formal study the evaluator should analyse the logical reasoning that connected programme inputs to desired outcomes to see whether there is a reasonable likelihood that goals can be achieved (Weiss 1997).

We developed a theory-of-change model during the scoping phase which should have provided a 'road-map' for the evaluation and for the pilots. This theoretical approach has been successfully used by many social researchers who are unable to implement rigorous experimental paradigms, but it appears that the BHLP pilots were not able to benefit from it, leaving the evaluation team with considerable challenges. Ideally, there should have been a period for reflection by the pilots after the findings from the scoping phase had been presented, but most were steaming ahead, primarily to spend the DCSF pump-priming funds by the end of the financial year. Thus, the lessons from our scoping exercise were not heeded.

4. Issues of study design

The degree of standardisation of an intervention is a key issue for evaluation of complex interventions. As is often the case in the evaluation of initiatives in health policy, a high level of control over intervention delivery is difficult to achieve in social care. Local delivery structures are a product of historic development, demographic factors and political direction. As in a pragmatic randomised controlled trial, standardisation may not be desirable even if it is possible. Many commentators have suggested that standardisation of intervention delivery will reduce effectiveness by impeding adaptation to local environments, which is possible in LA service provision. Hawe and colleagues argue that this does not preclude conducting a trial of an intervention (Hawe, Shiell et al. 2004). They argue that the intervention may take different forms provided that the essential functions are the same.

We share the view of Hawe and colleagues that standardisation is not a requirement for a valid evaluation. The success of BHLP working is dependent on the freedom of lead professionals to tailor a package of care that is relevant to individual needs and consistent with local delivery structures. This is not a bar to evaluation. A mixed methods approach to evaluation can provide insight into which aspects of the intervention and the local environments are conducive to success. However, there are a number of essential elements for a robust quantitative evaluation – the identification of the study population and an appropriate control group, the availability of suitable outcome measures and a procedure to collect data on intervention and control cases. All of these have posed problems in the evaluation of BHLP.

4.1 Defining the study population

Defining the study population for a complex social care intervention is often complicated by loosely specified eligibility criteria. We were dependent on the professional judgment of the LPs and BHLPs as to whether a child had 'additional needs', and the ability of local agencies to identify appropriate children. The ultimately criterion for intervention is that a child is likely to fail to meet one of the five outcomes specified in the Every Child Matters agenda. In practice, LPs and BHLPs are likely to respond to immediately identifiable needs, which may or may not correlate with these longer term outcomes.

The LAs all had existing mechanisms for multi-agency responses: some used a panel approach to discuss a number of children referred to one or more of the agencies; others had introduced the team-around-the-child in which a LP convenes representatives from other key agencies in order to discuss each case on its merits. Both these approaches mean that no one agency holds all the data about an individual child and, in most pilots, the BHLP co-ordinator could not provide us with co-ordinated data about the children with additional needs who had either a LP or a BHLP working with them. Inevitably, because the target group of children fell below the normal thresholds for statutory intervention, we could not obtain a list of all the children in each pilot area who would meet the criteria for either LP or BHLP intervention. Moreover, not all LAs had implemented the required CAF assessments and some continue to undertake local assessments which may not cover all the information gathered by a CAF assessment.

Although information-sharing protocols are in place in LAs, there are not common/shared data-management systems in place yet. This has produced another challenge both for the pilots and for the evaluation. Whilst coordination of services has moved towards a multi-agency framework, individual agencies retain control over staff. Records on service access are kept at home agencies on incompatible databases. Consequently the cost of data extraction for BHLP managers is extremely high. There are moves towards a unified system of data management alongside the introduction of an electronic common assessment form (e-CAF). Establishing the e-CAF as the only gateway to all targeted service provision along with a single database containing comprehensive details of service involvement would allow researchers to identify the study population with greater accuracy.

4.2 Identifying an appropriate comparison group

The selection of appropriate comparison groups is challenging whenever randomisation is not undertaken. Wolff suggests that there are limits to the applicability of randomisation in complex social care interventions (Wolff 2000), but a paper describing three examples of recent social care trials, illustrates how randomisation can be implemented in a complex social care setting (Oakley, Strange et al. 2003). The ethical case for randomisation is well established in health care (Hafner and van der Heiden 1991; Williams 1992) but dissent within social care still exists (Rose 1994). This dissent is

buttressed by concerns over equality of access to services and a failure to understand the purpose of randomisation, which can only ethically be used when there is no evidence that one intervention is superior to the other. The latter point partly reflects the prevailing concerns about evaluators in social care services that were touched on in section 3. . It may also reflect the perception that BHLP working allowed access to additional funds and therefore was clearly advantageous. We will examine the practicality of randomisation here.

In the BHLP study there are four discernible levels at which randomisation might have been implemented. Random assignment of a subgroup of the applicant LAs to the control is clearly feasible, but given the small numbers of LAs, would necessitate statistical adjustment for socio-demographic factors. Stratified randomisation would reduce the need for such baseline adjustment and is simple to implement. Randomisation of geographic areas within LAs is dependent on decentralised delivery structures more often found in geographically large, rural LAs. It would have been feasible in each LA except one densely populated urban authority that used a single assessment panel for the entire authority. Arguably a larger cluster size for randomisation might reduce the potential for demoralisation of families or professionals who were not assigned to BHLP, and contamination between groups. Randomisation of areas within each LA would have allowed an assessment of the effectiveness of BHLP working that was context specific. The comparison of effectiveness results between LAs would have helped identify LA characteristics and mechanisms which facilitate effective delivery of the intervention.

Randomisation of professionals to BHLP or LP working could have been implemented in the form of training a selected group of practitioners. However there is a risk of selection bias in the assignment of caseloads to professionals. Randomisation at individual or family level would ensure comparability of children receiving BHLP and LP care, but is likely to encounter resistance from professionals on the grounds of its apparent arbitrariness and unfairness. Inappropriate responses might include attempts at subversion and inhibition of appropriate financial caution.

Where multi-agency practice was well-established a borough wide BHLP scheme was generally implemented from inception. Most other LAs intended to roll-out the scheme across their boroughs as quickly as training and infrastructure requirements would allow. The importance of a comparison group was not made clear in the pilot tender documents as the evaluation team had been involved at the stage the DfES recruited the pilots. Also, it seems that LAs have not been discouraged to roll-out the initiative prior to the evaluation findings being known and, it is possible that they have been actively encouraged to move ahead by the external consultants providing a challenge and support role. Currently we are negotiating with LAs who still have a policy off area to try to persuade them to collect the data we need for a comparison with BHLP working.

The importance of suitable comparison groups in any policy evaluation must be recognised and incorporated into policy development. Tender guidelines

ought to make this a clear pre-requisite for a successful bid. A pre-trial evaluation phase that included scrutiny and suggested modifications to the schemes LAs had submitted might have allowed the DfES or the evaluators to establish appropriate comparators at commencement.

4.3 Identification of appropriate outcome measures

The long term goals of intervention have been clearly defined in the Every Child Matters agenda, but it is not clear which, if any, outcome measures provide a valid measure of progress towards these goals. This may contribute to the suspicion which any measures, and their application, are met from social care professionals. Complex initiatives, such as BHLP, where a diverse range of needs are targeted with broad long term goals, clearly require a range of outcome measures to try to fully capture changes following the intervention. This is especially important given the wide-ranging BHLP target group (children and young people aged 0-19). Our selection was severely constrained by the lack of institutional capacity at the LA level to collect any data. We have chosen the Strengths and Difficulties Questionnaire (SDQ http://www.sdqinfo.com/) which is a behavioural screening questionnaire, along with school absences for 5 to 16 year olds and Not in Employment, Education or Training (NEET) status for 16 to 19 year olds. None of these measures, however will assess the impact of interventions for families with very young children.

The SDQ was selected in part for its brevity, and partly because it is already widely used. Despite this we met considerable resistance from operational managers in some LAs, especially when asked to collect the data at two points in time. Whilst most LAs were committed to some form of evaluation this generally consisted of an assessment by the LP of whether the intended outcomes of the intervention had been achieved. A failure to appreciate the potential magnitude of bias in such a subjective assessment reflects the lack of a quantitative evaluation culture in LAs. The poor development of outcome measures for social care interventions has allowed an attitude that quantitative measurement is inappropriate, and even unethical, to remain unchallenged. The development of appropriate outcome measures was beyond the remit of this study but is clearly an area that needs further research if we are to be serious about assessing effectiveness and ‘cost effectiveness’ of social policy.

4.4 Data collection

The collection of data is central to any evaluation. Ideally data collection on resource use and outcomes would have been the responsibility of the evaluation team, but we were not resourced to do this. The evaluation of BHLP was dependent on the enthusiasm and resources of the BHLP co-ordinators to collect the primary data. Access to funding could be, and sometimes was, used to incentivize BHLPs to return data, but no leverage over LPs was available. This situation arose partly from the position of operational managers of BHLP within the organisational structure of each LA.

BHLP coordination is generally located within multi-agency teams which have no authority over individual agency staff.

The issue of data collection should have been addressed during a pre-trial evaluation phase. In all but one LA no provision at all was made for the collection of data as an explicit part of the proposed budget. All the pilots were aware that they would be required to provide data for a national evaluation, and all were sure they could achieve this. In practice, they have been quite unprepared for the demands of a quantitative evaluation. The importance of accurate and unbiased data collection is recognised in health care trials, where trial monitors are often employed. The potential for incomplete, inconsistent and inaccurate reporting appears to be just as great in evaluation of social policy. Ideally, Identification and resourcing of data collection should have been established before commencement of the pilot, with the Government arbitrating between the needs of evaluators and the practical difficulties of LA implementers prior to acknowledging the proposed LA budget spending plans.

5. Issues specific to economic evaluation

Before the identification of relevant cost inputs can take place, it is important to define the perspective of the study. A societal perspective requires that all resources used to provide interventions and all future resources saved by the success of the intervention should be included (Fox-Rushby and Cairns 2005). In reality the identification of all the costs and savings accruing to families and the wider community over the long term is likely to be a mammoth task. Hence this study restricted consideration to those costs and savings falling on local authority budgets.

5.1 Costs

Complex initiatives delivered in heterogeneous environments provide extra challenges for evaluators to cost; it may be beneficial for evaluators to spend further time identifying the main components of the programme within each pilot site to facilitate costing. The identification of all resource inputs can be challenging for complex interventions where the mode of action is difficult to deconstruct (Hawe, Shiell et al. 2004). Thus it is possible that evaluators may not be able to explicitly identify and measure all inputs. In addition, there are far less nationally accepted unit costs available for both statutory and non-statutory services that act as inputs to wider public sector services outside health (Byford and Sefton 2003). This inevitably leads to less accurate cost data.

There are a number of methods available to measure the quantity of resources used by study participants, including questionnaires, diaries or case notes (Johnston, Buxton et al. 1999). Complexity of interventions has an impact on the use of these data collection methods (Byford and Sefton 2003). Economic questionnaires, completed by researchers or self reported, are

often used for evaluation purposes. The scale of data collection often dictates the use of self reported forms, although this can reduce the accuracy of data. The large numbers of agencies and inputs required for complex interventions complicates data collection with service diaries, and there may be problems with recall accuracy. Case notes are another avenue that economists have traditionally used particularly for resource use in the biomedical research. Again there may be additional data collection burdens if multiple agencies use different recording systems.

During the course of the BHLP scoping phase, it became apparent that resource use was wide and varied across multiple professionals and agencies, and there were no central records kept on a case basis. This is not unusual with multi-agency working, there are many obstacles hindering data collection including information sharing and adequate IT systems. The lack of common data systems within the BHLP pilot sites favoured the use of self reported health economics questionnaires but resistance from LAs impeded their use. LAs have no central data on the cost of services provided by the agencies they coordinate. Consequently we have to rely on very crude estimates of service costs.

It is quite possible that episodic cost data is available but is fragmented over the myriad data systems that various agencies use. Again, a central IT management system and subsequent information sharing might facilitate more appropriate commissioning by LAs as well as providing accurate cost data for evaluators. Publication of national reference costs for social care interventions similar to those published for health by the Personal and Social Services Research Unit (http://www.pssru.ac.uk/) would be invaluable.

Identification of and data collection on cost savings in any evaluation, usually requires long term follow up. The future effects of BHLP intervention on a child may be felt over a very long time frame. The estimation of these impacts is challenging due to the lack of evidence base regarding the long term impacts of social care interventions. Consequently the BHLP evaluation will not be able to consider long term effects.

5.2 Effects

In health care evaluations we are usually able to choose between evaluation types based on use of a narrower or broader (and often utility-based) outcome measure. There is a lack of a broader measure of quality of life, analogous to the quality adjusted life year, which addresses the burden of social problems. Consequently, we are unable to attach a utility weight to the effects of social care interventions. We could attempt to measure the willingness to pay for social care interventions, but it is not clear who we should be asking for this evaluation – the recipients or those paying for the intervention.

A cost-effectiveness analysis, defined as the comparison of incremental opportunity costs with incremental non-monetary outcomes which are common to both interventions, avoids this dilemma (Drummond, Sculpher et al.; Garber and Phelps 1997). However, complex interventions have multiple

effects requiring a range of outcomes. There have been efforts to overcome these problems and take multiple outcomes into account (Bjorner and Keiding 2004), but this can be difficult (French and Drummond 2005).

We chose to evaluate the BHLP initiative using the SDQ, school attendance (5-16s) and NEET status (16-19s). Although the SDQ score is the main outcome measure to be utilised in the evaluation it is evidently insufficient to capture the effects of BHLP in all cases. Where no one dominant outcome measure exists for all cases in the study, a more disaggregated approach, often labelled 'cost consequences analysis', is often undertaken. However, such analyses have their shortcomings in that although they involve the presentation of a range of outcomes and costs, no attempt is made to combine them. As such, decision makers are left to form their own opinions which may or may not be the most efficient choice.

6. Where now?

Two key areas of difficulty emerged from our scoping evaluation of the BHLP initiatives that the sixteen chosen LAs had implemented. Primarily, it was evident that we would have severe problems accessing the data we needed. The collection of resource use and outcome data was the responsibility of the BHLP coordinators, but their institutional capacity to do this was highly limited. Our choice of outcome measures in particular was curtailed by these limitations. The scoping evaluation also uncovered a disparity between the policy intent and practice in the sixteen LAs. To a greater or lesser extent each of the LAs had implemented the BHLP initiative as a 'top-up fund' providing grant funding for additional goods and services that could not be accessed through the normal routes. Finally the lack of an appropriate comparison group in most LAs has restricted the application of a comparative evaluation.

Given the difficulties outlined in this paper we have had to take a flexible approach to the evaluation. Where we can identify and collect data on an appropriate comparison group we can apply a more rigorous model which will compare BHLP working with LP working. In the absence of a comparison group we suggest a more limited evaluation which attempts to compare the cost-effectiveness of goods and services provided by the BHLP budget with traditional service interventions. This limited comparison treats BHLP as an individual fund source rather then an alternative approach to service provision, and while it may reflect what is being delivered in many LA, it may not capture all of the effects of BHLP working.

In LAs where some areas used BHLP working and some areas used LP working we can apply a robust model which compares the effectiveness and cost-effectiveness of BHLP and LP working. Both the primary outcome measures (SDQ) and the secondary outcome measures (school absences; NEET status) could be analysed. The model is outlined below: (***Bold + italics*** refer to the parameter of interest in each regression.)

i. SDQ(review) ~ SDQ(assessment) + age + gender + Ethnicity + ***Area(BHLP or LP)***

If the effect of ***Area(BHLP or LP)*** is statistically significant it implies that outcomes differ significantly for children receiving BHLP and LP care; if the coefficient of ***Area(BHLP or LP)*** is positive, it implies that children receiving BHLP care are better off; if it is negative, it implies that children receiving LP care are better off.

ii. SDQ(review) ~ SDQ(assessment) + age + gender + Ethnicity
+ ***Area(BHLP or LP) + expenditure***

If the coefficient of ***Area(BHLP or LP)*** is statistically significant in model (i) but non-significant after allowing for expenditure in model (ii), this implies that children respond better in one area only because they have had more money spent on them.

iii. SDQ(review) ~ SDQ(assessment) + age + gender + Ethnicity + Area(BHLP or LP)
+ LP expenditure + BHLP expenditure

The coefficients of the terms ***LP expenditure*** and ***BHLP expenditure*** indicate the cost-effectiveness of LP and BHLP working respectively. If the coefficient of BHLP expenditure is significantly greater than the coefficient of LP expenditure this implies that children respond better for the same amount of money if the money is from BHLP care rather than LP care.

There are concerns with the application of this model. Only the most basic demographic information is included as co-variates, and we cannot be certain that the two groups are comparable in the absence of randomisation. A second fear is that the outcome measures we are using are better suited to measuring progress from traditional service interventions than from non-traditional services and goods bought from the BHLP budget.

In the majority of LAs, borough wide roll-out of BHLP will prevent the identification of any LP comparator. The evaluation will be limited to a comparison of the effectiveness of goods and services purchased with the BHLP budget and those normally provided by the LA and voluntary sectors. Changes in the three chosen outcome measures (SDQ, school attendance and NEET status) per pound of service expenditure from the BHLP budget would be compared with that achieved per pound of expenditure on traditional services. The proposed model for comparison is essentially simple. Linear regression will be performed using a model of the form:

SDQ(review) ~ SDQ(assessment) + age + gender + Ethnicity
+ BHLP expenditure + LP expenditure$|_{\text{BHLP children}}$ + LP expenditure$|_{\text{non-BHLP children}}$

LP expenditure among children who do and do not receive BHLP will be considered separately, as the effect of LP spending may be different among children receiving BHLP interventions and those not. We will test whether the cost-effectiveness of LP expenditure among children who do and do not get BHLP is similar; if it is, we will combine these terms.

This model will allow assessment of whether:

- children respond better if they have more (BHLP) money spent on them (after allowing for baseline differences between children)
- children respond better if they receive more money spent on traditional interventions (after allowing for baseline differences between children)
- BHLP expenditure is more cost-effective (produces a larger change in SDQ per pound spent) than LP expenditure

There are a number of additional concerns with this approach. Whilst all children in LAs with a borough wide scheme may be receiving care from a BHLP, those receiving additional expenditure from the BHLP budget clearly have a spectrum of needs not met by traditional services. These children may be systematically more or less responsive to service provision in general than other children. Examining the coefficients of $\text{LP expenditure}|_{\text{BHLP children}}$ and $\text{LP expenditure}|_{\text{non-BHLP children}}$ will give us an insight into this problem, but if the coefficients are significantly different we do not know whether this is due to differences in the children selected or differences in the service mix they received.

The risk of selection bias in comparing the outcomes of children selected by professionals to receive any particular intervention is obvious. We would expect professionals to choose children likely to respond well to the proposed intervention, arguably this is an integral part of the intervention. That these children might also respond better to the package of traditional services arranged seems plausible. Hence there are dangers in concluding from this comparison that BHLP is cost-effective compared to LP working. It seems less likely that the direction of bias would run the other way. Hence this comparison might be considered a preliminary hurdle rather than a definitive measure of cost-effectiveness.

7. Conclusions

In this paper we have outlined several challenges to undertaking comprehensive economic evaluations in the field of social care and complex public health interventions. There is an extent to which economic evaluation (usually labelled ‘cost effectiveness’) is ‘tagged onto’ pilot study briefs and are, thus, not adequately planned in advance or resourced.

We see two main sets of challenges; one of culture and one related to meeting more practical considerations related to undertaking such studies.

With respect to culture, the challenges seem to be:

- establishing the perception of evaluations as assessing the merits of a programme rather than simply ironing out teething troubles;
- overcoming fundamental misunderstandings and misgivings about the nature and value of quantitative evaluation which contribute to an inability to collect data and decisions over mainstreaming the initiative that are not supported by any quantitative analysis.
- for Government to invest in clearer policy articulation, but especially putting across how this relates to the evaluation;
- trying not to implement changes too fast and thus, in ways that actually hamper the original policy intent;
- for researchers to be involved very early, at the pre-evaluation phase – starting at tender where Government, LAs and evaluators work together to establish evaluable schemes in each LA, establishing notions such as selective roll-out to (randomised) recipients. (Of course, this recommendation is made difficult by Government tendering processes whereby the evaluators are not decided upon until the pilots schemes have commenced.)

With respect to the actual conduct of evaluations, the challenges are:

- agreeing who is best placed to ensure data quality (accuracy, consistency and completeness), which also requires a discussion about resourcing of this activity along with appropriate incentives;
- improvement in infrastructure which are required to build a common multi-agency database, appropriate records and care episode costs.
- more appropriate outcome measures: this is likely to require extensive consultation with professionals, families and children to ensure that outcome measures are acceptable and appropriate, and systematic review of the academic literature to identify the availability and validity of current instruments.

We recognise that the specific evaluation we were invited to conduct is being undertaken at a time of significant changes in children's services, with provision shifting to a professional-led style of working (with budgets). A longer-term ongoing evaluation might still yield important policy results. We have also outlined what we see as a fairly flexible approach to a cost effectiveness analysis of BHLP which may have broader applicability. However, it is evident that much more work needs to be done here, both to address the limitations of the analysis we propose and to further develop measures of effectiveness for social care interventions. Given that we have still to implement this model, we welcome comments both specific to our own project and also to the issues we have raised in general.

References

Bjorner, J. and H. Keiding (2004). "Cost-effectiveness with multiple outcomes." Health Economics **13**(12): 1181-90.

Byford, S. and T. Sefton (2003). "Economic evaluation of complex health and social care interventions." National Institute economic review(186): 98-108.

Campbell, M., R. Fitzpatrick, et al. (2000). "Framework for design and evaluation of complex interventions to improve health." BMJ **321**(7262): 694-6.

DfES (2004). "An overview is available at :http://publications.everychildmatters.gov.uk/eOrderingDownload/1677-2005PDF-EN-01.pdf." Department for Education and Skills.

DfES (2006a). "Information sharing: Practitioners' guide http://publications.everychildmatters.gov.uk/eOrderingDownload/0338-2006BKT-EN.pdf." Department for Education and Skills.

DfES (2006b). "The Common Assessment Framework: Practitioners guide http://publications.everychildmatters.gov.uk/eOrderingDownload/DFES-0337-2006.pdf." Department for Education and Skills.

DfES (2006c). "The lead professional - Practitioners' guide http://publications.everychildmatters.gov.uk/eOrderingDownload/0335-2006BKT-EN.pdf." Department for Education and Skills.

DfES (2006d). "An overview is available at :http://www.everychildmatters.gov.uk/deliveringservices/leadprofessional/budgetholding/."

DoH (2004). "http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_4094550." Choosing Health, Department of Health.

Drummond, M. F., M. J. Sculpher, et al. "Methods for the economic evaluation of health care programmes." 379p.

Fox-Rushby, J. A. and J. Cairns (2005). Economic evaluation. Maidenhead ;; New York, Open University Press.

French, M. T. and M. Drummond (2005). "A research agenda for economic evaluation of substance abuse services." Journal of Substance Abuse Treatment **29**(2): 125-37.

Garber, A. M. and C. E. Phelps (1997). "Economic foundations of cost-effectiveness analysis." Journal of Health Economics **16**(1): 1-31.

Hafner, H. and W. van der Heiden (1991). "Evaluating effectiveness and cost of community care for schizophrenic patients." Schizophrenia Bulletin **17**(3): 441-51.

Hawe, P., A. Shiell, et al. (2004). "Complex interventions: how "out of control" can a randomised controlled trial be?" BMJ **328**(7455): 1561-3.

Hawe, P., A. Shiell, et al. (2004). "Methods for exploring implementation variation and local context within a cluster randomised community intervention trial.[see comment]." Journal of Epidemiology & Community Health **58**(9): 788-93.

Johnston, K., M. J. Buxton, et al. (1999). "Assessing the costs of healthcare technologies in clinical trials." Health Technology Assessment (Winchester, England) **3**(6): 1-76.

Judge, K. and L. Bauld (2001). "Strong theory, flexible methods: evaluating complex community-based initiatives." Critical Public Health **11**(1): 19-38.

Kelly, M. P. (2005). "Economic appraisal of public health interventions." London : Health Development Agency.

Kelly, M. P. M., David; Ludbrook, Anne; Powell, Jane (2005). "Economic appraisal of public health interventions." NHS Health Development

agency, available at http://www.chsrf.ca/kte_docs/Economic_appraisal_of_public_health_interventions%5B2%5D.pdf.

Lafferty, C. K. and C. A. Mahoney (2003). "A framework for evaluating comprehensive community initiatives." Health Promotion Practice **4**(1): 31-44.

MRC (2000). "A framework for the development and evaluation of RCTs for complex interventions to improve health." Medical Research Council.

Oakley, A., V. Strange, et al. (2003). "Using random allocation to evaluate social interventions: three recent U.K. examples." Annals of the American Academy of Political and Social Science **589**: 170-189.

Rose, H. (1994). "Love, power and knowledge." Cambridge: Polity.

Shemilt, I., I. Harvey, et al. (2004). "A national evaluation of school breakfast clubs: evidence from a cluster randomized controlled trial and an observational analysis." Child: Care, Health & Development **30**(5): 413-27.

Susser, M. (1995). "The tribulations of trials--intervention in communities." American Journal of Public Health **85**(2): 156-8.

Weiss, C. (1997). "Theory-based evaluation: past, present and future." New Directions for Evaluation **76**: 41-55.

Williams, A. (1992). "Cost-effectiveness analysis: is it ethical?" Journal of Medical Ethics **18**(1): 7-11.

Wolff, N. (2000). "Using Randomized Controlled Trials to Evaluate Socially Complex Services: Problems, Challenges and Recommendations." The Journal of Mental Health Policy and Economics **3**: 97-109.