

Evaluating Evaluations

Evaluating recent evaluations of Sure Start, Home-Start and Primary Age Learning Study

Helen Barrett

RESEARCH & POLICY FOR THE REAL WORLD

Contents

PREFACE	1
Introduction	2
Three Early Intervention Programmes: Home-Start, PALS and Sure Start	3
HOME-START	4
The Belfast Study	5
The London (Birkbeck) Study	6
Conclusion	8
PALS - PRIMARY AGE LEARNING STUDY	9
SURE START	10
The National Evaluation of Sure Start	11
The local context analysis	11
Preliminary findings: local context effects	11
SSLP Implementation	12
Preliminary findings: implementation	12
SSLP Cost-effectiveness	13
Support for SSLP self-evaluations	13
Preliminary findings: impact	14
DISCUSSION	16
Problems for evaluation: what are we measuring?	16
NESS: The need for caution in drawing potentially premature conclusions from early findings	17
How widely applicable are 'intention to treat' designs?	17
Are we asking the right questions?	18
The need for evaluations more directly tailored to capture effects of interventions on individual families	18
CONCLUSIONS	19
REFERENCES	20

The Family and Parenting Institute is the leading centre of expertise on families and parenting in the UK. Families, in all their diversity, form the basis of our society and the foundation for the future. Our mission is to support families in bringing up children.

Visit our website at

www.familyandparenting.org for more information about our work and our other publications.

About the author

Helen Barrett has researched on and worked with children, their families and their carers in a wide range of family and community settings. Formerly a senior lecturer in developmental psychology, she joined the Family and Parenting Institute in 2002. She is particularly interested in close attachments and the influence of non-traditional care settings and separation experiences on children's and parents' emotional development and relationships.

© Family and Parenting Institute 2007

Published by
The Family and Parenting Institute
430 Highgate Studios
53-79 Highgate Road
London NW5 1TL
Tel 020 7424 3460

Email: info@familyandparenting.org
www.familyandparenting.org

ISBN 1 903615 48 8
978 1903615 48 2

Registered charity no.1077444

Preface

Evaluating Evaluations is the review written for the seminar: *Commission in Haste, Repent at Leisure? Evaluations of Family Preventative Services and the Implications for the Development of Policy*. This day-long seminar, held at the House of Commons in April 2006 and commissioned by the Family and Parenting Institute and the Joseph Rowntree Foundation, drew together an audience of some 100 policy makers and academics to discuss the lessons that have emerged from family support evaluations. A lively debate took place reflecting the increasing emphasis on using evaluative research to inform government decision making in relation to supporting families and enhancing outcomes for children. Recent disappointing evaluation results prompted a number of questions, for example whether the evaluative bar is being set too high in relation to the impact of preventative services on families.

A report was written for the seminar by Dr Helen Barrett, Senior Research Fellow at the Family and Parenting Institute, analysing major evaluations of family support, specifically evaluations of Sure Start, Home-Start and the PALS programme. The review included discussion of the methodology of evaluations and the nature of the findings, their validity and implications for the purposes of policy influence. It offered a highly informative analysis of family service assessment, and is published here in order to reach a wider audience of policy makers, academics and professionals for whom evaluation is a critical factor in developing a new and often contentious service field.

Clem Henricson

Director of Research and Policy
Family and Parenting Institute

Introduction

There has been considerable media attention paid to recent evaluation studies of two UK early intervention programmes, Sure Start and Home-Start. Preliminary findings in relation to Sure Start gave little evidence of the positive effects that the programme had been expected to achieve, and were seized upon by some as an opportunity to question the efficacy of New Labour's ambitious efforts to reduce child poverty and to lower the risk of social exclusion. Rather than entering into this more political debate, the intention here is to examine the evaluation reports, their claims and methodology, and to make as objective as possible an assessment of the significance of their findings. In doing this, skills of critical analysis acquired through training and experience in the use of both qualitative and quantitative approaches to data collection and analysis are drawn upon, though the latter is the primary focus due to a conviction that it is only through these that causal relationships can be examined.

While recognising the difficulties involved in setting up randomised controlled trials (RCTs), it is generally accepted that these are the 'gold standard' to be aimed for. Not only do they ensure that like is compared with like, that is, that as pure-as-possible a 'no intervention' condition is compared with as pure-as-possible an 'intervention' condition, but they also enable effects of interventions to be isolated so that we can be more confident in concluding that specific outcomes are related to the intervention rather than to extraneous or irrelevant influences. They therefore have a number of important features that other designs lack:

- controlled manipulation of cause and effect
- exclusion or control of irrelevant or extraneous factors
- identical sub-samples in intervention (study) and non-intervention (comparison) conditions
- study and comparison groups both allocated to condition randomly
- judgements of outcome made by researchers blind to condition, that is, by people who are not in a position to tell whether participants have received the intervention or not.

Designs that lack any one of these features can make it difficult to tell whether outcomes are true effects of interventions or whether they might actually be due to pre-existing differences between study and comparison groups, or direct effects of other influences such as response or observer biases.

Many people are sceptical of or impatient with research approaches that take strictly empiricist or experimentally-oriented approaches to social scientific investigation. Some

consider them unrealistic, old-fashioned or perhaps better reserved for the physical sciences, and some argue that they are not capable of capturing the true complexity of human social and psychological development. People taking these views tend to stress how children in the UK grow up in family contexts that are diverse, multi-cultural, and constantly changing, and argue that different kinds of methodology are needed to capture these moving targets. Others, quite rightly, point out the extent to which even the most rigorous scientific approaches can be distorted by unintended pre-conceptions and biases.

Even more serious opposition comes from other sources because, historically, in the UK, there has been a tendency for social research to be seen as a soft and rather inadequate alternative to 'real' science. Perhaps partly because of this attitude, funding for social scientific research has been relatively hard to come by and is often inadequate for the tasks set. Under the circumstances, the temptation to be less than rigorous can sometimes become a requirement for survival. Unrealistic demands made by funding bodies evoke unrealistic promises from those seeking funding. Short, quick, corner-cutting methods feature as much if not more prominently than more rigorous methodologies and findings from opinion polls. The latter, which are frequently based on non-representative samples, or samples whose provenance is unreported, can receive as much if not more attention in the media. As a result, tendencies to contest, doubt or disrespect findings from research in the social sciences are encouraged and the considerable skills involved in sound evaluations can be de-valued. The scene is also set for policy-makers to cherry-pick research findings, selecting only those that support policies and practices that are currently in vogue.

There is no doubt that evaluating early interventions is not an easy task. The development of any single individual is immensely complex and that of groups of individuals is even more difficult to track. Countless studies have underscored the extent to which development is in most aspects non-linear and multiply determined; and that it has multiple goals, which can change, and multiple ways of reaching those goals. Models that do not reflect this are likely to be based on false premises and, therefore, are more likely to lead to contestable conclusions.

It follows, then, that when an early intervention programme is introduced, its effects cannot be assumed to lead to the same outcome for all those in receipt of it. It is not like dripping bright red paint onto a sheet of slightly damp, blank paper and watching as a rosy glow emerges. We might have bright red paint but the paper is neither blank, nor uniformly

wet, nor of a uniform texture. Any appraisal of effects is also likely to be influenced by the quality of the lens through which they are viewed and the way they are looked at. For all these reasons, it is unlikely that simplistic models of human development will help us to understand the effects of intervention programmes on outcomes for any individual or group of individuals. Rather, the task of evaluating effects of early interventions must necessarily be highly complex because so many potential influences need to be taken into consideration.

This being the case, it also follows that evaluating evaluations of early interventions can hardly be any less tricky. This involves all the same processes but at one remove. Crucially, in taking on such a task, we must rely on being given access through evaluation reports to all the requisite details that permit a fair assessment to be made. This means comprehensive details both about the design of the study and about the design of the evaluation, details such as how samples were obtained, who they involved, how appropriate comparisons were, how data was collected and treated, and how it was analysed. Such basic details, too often, are not supplied in evaluation reports. Without them, credence cannot be given to the claims that evaluators make, not necessarily because they have carried out poor evaluations but simply because it is not possible to tell from the evidence that they have presented whether or not they have done a good job. In other words, although claims can always be made at any stage in the evaluation of an intervention, there is a clear distinction to be made between claims based on thorough evaluations and those based on more slender evidence. From the point of view of the press and public, all claims may sound equally plausible. From the point of view of social scientists and of policy-makers, only those claims based on sound methodology should count.

So, what claims have been made by the authors of the reports under consideration here? Reports from the press (e.g. *Daily Mail*, *Guardian*, *Sunday Times*, September 2005) would suggest that the main finding from evaluations both of Sure Start and Home-Start was that a lot of money has been wasted on schemes that make little difference to the children and families they were intended to benefit. The preliminary findings of the PALS trial, by contrast, appear to have indicated that parenting practices can be improved even if a substantial number of parents targeted have not taken part in the intervention or have attended less than a third of available sessions and even though this programme did not show any direct benefits in terms of children's reading ability or reduction of antisocial behaviour. What are we to make of this?

The task undertaken here is to present more fully the findings from the evaluation reports, so that a fair appraisal can be made of what they actually showed as opposed to what the press or perhaps even the authors claim to have shown. First, each of the three interventions is briefly described, then the evidence so far accumulated in respect of their effectiveness is examined. In doing this, the focus is more upon the National Evaluation of Sure Start (NESS) than on the other two intervention evaluations because, in many respects, this demonstrates rather fully the issues that seem most central to a debate about policy and practice around early intervention programmes. However, the evaluations of the other two interventions will be touched on and, from consideration of all three, a summary of their major findings and of the issues associated with interpreting these findings will be extracted.

Three Early Intervention Programmes: Home-Start, Primary Age Learning Study (PALS) and Sure Start

The three programmes share three essential aspects in common:

- First, they all aim to change outcomes for children growing up in family environments where there is a high risk of significantly less than optimal development, that is, families affected by multiple sources of social and personal disadvantage, such as poverty, mental ill health, domestic violence, criminality etc..
- Second, they aim to do this by influencing the course of early development. The intention here is to ensure that the child's developmental trajectory is fundamentally altered so that the child is set onto a more optimal route or track.
- Third, they each do this by influencing the child either directly or indirectly, that is, by changing the way the child is being cared for or the conditions within which the child is growing up.

These aspects are based on arguments such as those that arose from the Glass review (Glass 1999), which suggested that early interventions are not only the most likely to be effective, because at this stage there is still scope for 're-wiring the human organism', but that they are also the most likely to make cost-effective changes, because they are proactive and preventative rather than reactive and crisis-driven. Such arguments have a great deal of intuitive appeal, are encouragingly full of optimism and are seen as almost self-evident. As a result, few dispute them and those who do tend to be swiftly dismissed. However, questions still need to be asked about the true degree to which we can predict or control developmental destinies.

This fundamental philosophy apart, in other respects, Sure Start, Home Start and PALS are not so similar. Their methods (though there is some degree of overlap) are different as are the points at which they engage the population of people deemed to be at risk. The three evaluation reports also reflect considerable divergence in strategies for evaluating respective programmes. These broad differences are summarised in Table 1.

As stated above, Home-Start aims to strengthen the support network of vulnerable families, on the understanding that good social support is one of the main protective factors capable of promoting resilience under stressful living conditions. Logically, therefore, an assessment of the effectiveness of Home-Start would require, as a first step, that measures should be taken of the existence of support in the social network of families identified as vulnerable, preferably prior to any intervention or offer of intervention. This step poses a major problem for evaluation since it has

Table 1: Characteristics of the three interventions

	SURE START	HOME START	PALS
Population of interest	Primarily 0-4, + pregnant mothers	Families around the time of birth	4-5 year olds and their parents/carers
Recruitment	Geographical areas – indexed as low on resources/services, high on social disadvantage/risk	Referrals (health services, other professionals or paraprofessionals, or self)	Reception and Year 1 classes in areas of social disadvantage
Intervention type	Universal/Targeted	Targeted	Mostly targeted
Method	Various	Befriending	Standardised
Length of intervention	Variable	Variable, up to school age	16 half-day sessions
What is changed?	Access to quality services; Attitudes to child care; Attitudes to health care; Childcare facilities	Social support network; Mother's emotional state; Ability to care for child	Parenting skills; Parents' involvement in teaching children to read; Reading readiness

HOME-START

Home-Start aims to strengthen and widen the social support network of families, focusing usually on mothers, around the time of birth, through trained volunteers or 'community mothers' who befriend the vulnerable family. Referrals, usually from health visitors, are made on a variety of social, physical or psychological grounds. Befrienders, who are usually parents themselves, visit each family in its own home. They offer support, friendship, reassurance and practical help, such as encouraging effective use of services within the local community.

Two recent evaluations of Home-Start, the first carried out by McAuley and colleagues at Belfast University (McAuley 1999, McAuley et al. 2004, Slead et al. 2005, McAuley et al. 2006) and the second by Barnes and colleagues at London University (Barnes et al., *in preparation*), did not find advantages for programme recipients and, in the case of the Barnes et al. study, indicated a negative association between mothers' participation in the programme and cognitive development in children.

proved to be very difficult and time-consuming to measure actual social support.

The preferred solution for many researchers has therefore been to measure perceived support, on the assumption that changes in perception of support reflect changes in actual support. Clearly, this is not a safe assumption since perception of support can fluctuate considerably, depending upon a range of individual attributes or internal factors (e.g. mood, self esteem, depression, anxiety, personality, attribution style) as well as external factors (e.g. freedom from family demands, physical mobility, availability of friends or family members, advice, practical help). While some might argue that perceived social support can provide a more accurate indication of emotional wellbeing than actual support, there are good reasons to doubt this assumption. For example, socially phobic isolates, with almost no actual support, may rate themselves as having more than enough support, while highly gregarious individuals, with many close friends, may rate themselves as lacking sufficient support. Similarly, people experiencing manic-depressive emotional patterns may, in a downswing, rate themselves as low on support but, in another phase and with no change in actual support, may

rate themselves as high on support. These examples illustrate the difficulty of taking accurate measures of changes in the quality or quantity of social support.

Aside from this problem, two other aspects of Home-Start make for further difficulties in respect of evaluation. First, making any substantial change in the quality of a family's social support network is almost certainly likely to take time since it will often depend upon altering the quality of existing relationships or making new relationships. Both of these alterations need time to establish and time to test. Benefits are unlikely to appear as a short-term outcome. Second, families offered Home-Start are heterogeneous: some may be offered Home-Start because they have experienced a multiple birth and are perceived to be in greater than usual need for support for this reason; other mothers may need practical help around managing their own physical disability; yet others may have psychological and/or physical health problems, or have experienced difficult pregnancies or births, or have more complex needs. Given this diversity, it cannot be assumed that the same set of outcome measures will capture the nature of benefits accrued to individual families. Thus, there are at least three features of Home-Start that pose particular practical and theoretical challenges to researchers tasked with evaluation of its effectiveness.

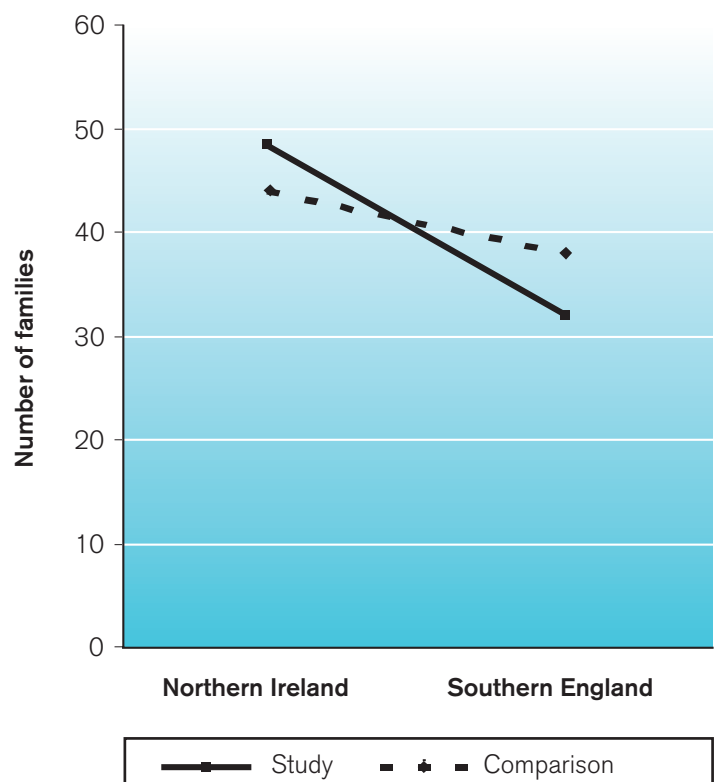
The Belfast Study

The Belfast study of Home-Start involved semi-structured interviews and administration of standardised tests (such as the Edinburgh Postnatal Depression Scale, the CES-D depression scale, the Parenting Stress Inventory, a shortened Rosenberg's Self-Esteem scale, the Maternal Social Support Inventory and a child development scale). Reports to date have neither presented full details of recruitment of participants nor details of outcomes relating to standardised measures. In the absence of this information, it is very difficult to make sense of the information that is given.

With regard to the sample, it would appear that 'pre-trial' interviews were carried out after the Home-Start intervention began and standardised measures were also taken at this time. Follow-up interviews and measures were taken approximately eleven months later. This raises questions about the timing of assessments, not only because eleven months may not have been long enough for changes to support networks to have taken effect, but also because the 'pre-trial' assessment actually took place after the intervention began.

No details of how the comparison group was recruited are reported. This makes it impossible to be confident that study and comparison group families were indeed equivalent in crucial respects. The information given about sample size suggests that there were 88 study group and 89 comparison group families in the study in May 2002, with 92 of these families residing in Northern Ireland and 70 in Southern England. Fifteen families had dropped out by the time of the second follow up, leaving 80 study group and 82 comparison group families. The information provided does not allow readers to ascertain the location of the families still in the study by this point. Though numbers are not given in the text, it would appear from the bar charts presented that, at the outset, unequal numbers of comparison and study group families resided in the two different geographical locations (see Figure 1).

Figure 1: Location of Belfast Home-Start study families



Information is not provided in the text about what the comparison group families were told about the study. These families took part in two fairly lengthy interviews and completed the same standardised tests as the study group families. They, like the study group families, completed the Client Service Receipt Inventory which elicited information about services used in the three month period prior to the

first assessment and during the months between first and second interviews. It is not clear whether comparison group families were aware that they were not being provided a service or what sense they made of their contribution to the study. From the information given, therefore, it is not possible to ascertain the extent to which response biases or researcher effects may have affected results.

Further details in the text suggest that referral practices, the nature of families, and their presenting problems and referral practices differed by geographical location. For example, the Northern Irish families were larger, higher on parenting stress and on clinical levels of stress, and there were more referrals for social isolation and maternal mental health problems in the Southern English sample. Other differences were also present such as higher numbers of multiple births in the Irish study group families. Given these initial disparities, it seems that the most appropriate statistical analysis may have been a mixed multivariate factor analysis, with two unrelated factors (country: Ireland/England; group: comparison/study) and one related factor (time of testing). Consideration would then need to have been given to the question of whether the sample size was large enough for such an analysis. Again, in view of the initial differences between families in each geographical location, it is not advisable to treat the data as though it were all drawn from the same population. This approach would confound location with study group status and would render it impossible to interpret findings of difference between groups with any degree of confidence: differences could either be due to participation in the programme or to geographical location, or both. There is no way of knowing.

The authors report neither multivariate nor univariate analyses of outcomes from interview data or standardised measures. Instead, they present results only of a subsample of 42 mothers selected as having an "average number of stress factors at the start" and assert that all mothers, regardless of whether they were recipients of Home-Start reported less stress approximately one year later, with the exception of stress due to finance. No benefits were found for the sub-sample of children assessed using the child development scale. However, given that the scale was suitable only for children under three, was only used in relation to one child in each family and its reliability had not been established, this seems unsurprising. Overall, they conclude from this that there is no evidence that Home-Start is of benefit to recipients or that it is a cost-effective programme.

Unfortunately, due to the lack of information supplied in published reports, it is inadvisable to place confidence in the

claims made by the authors of the Belfast Home-Start study. In time, perhaps more evidence will be presented to support the assertion that there are no benefits. However, on the basis of reports that do not supply sufficient information and a study design that appears to contain some conceptual weaknesses, it seems inadvisable to place confidence in preliminary findings which, at this time, appear to have doubtful evidential status. (This in no way precludes the possibility that fuller reports will later provide more substantial evidence).

The London (Birkbeck) Study

Reports from the London (Birkbeck) Home-Start study, though as yet unpublished, contain a rather more comprehensive account and so permit a more informed appraisal. It should be noted though that, for the purpose of the study, the Home-Start scheme was offered either proactively (just before or after the birth of a child) or, as it normally is, reactively to families assessed as being at risk. This means that the study cannot be said to be an evaluation of the Home-Start programme as it normally operates.

A further departure from Home-Start's usual practice was that families were recruited into the study on the basis of scores on a Social Disadvantage Index, an instrument devised to assess material aspects such as housing occupation, tenure, vehicle ownership, mother's education, neighbourhood resources, etc.. It is not known how the indication of vulnerability provided by these scores would relate to individual family vulnerability as assessed through more usual referral procedures.

Regardless of the point at which the intervention began, mothers were interviewed two months after the birth and again ten months later, when babies were approximately twelve months old. The design cannot therefore be said to be a pre-/post-intervention design and the two month measures may not have been true 'baseline' measures of mothers' emotional state prior to the offer of Home-Start. This initial complication seems likely to have made for a fundamental design problem which would make interpretation of results difficult.

The reason for choosing two months after birth as the first point of testing is not explained in the report. Although it perhaps seems sensible to test mothers and babies when babies are all at about the same chronological age, such an approach may not necessarily be the most appropriate when it is the potential impact of an intervention that is to be assessed. It is well established that the first two months after birth are a major period of transition in the lives of mothers,

particularly for mothers of first-borns (it is not clear whether mothers in the study were mothers of first or later born children). Given this fact, it seems likely that all mothers at this point are likely to have elevated levels of anxiety that may be expected to decline as they adjust to the new baby. We should consider, then, the possibility that these 'normal' yet quite distinct changes in emotional state could be expected to obscure changes due to other possibly weaker influences. Thinking about this problem in more concrete terms, we might imagine two groups of babies being observed for effects of growth hormone treatment. The design is equivalent to measuring babies' length at two months, when babies in the treatment group have received between a few days and two months treatment while babies in the no-treatment group have not received any growth hormone treatment, then measuring length again at twelve months. To make an accurate assessment of the effects of growth hormone treatment, it may not be enough just to compare overall rates of growth between the two groups of babies after ten months' treatment (give or take two months due to the staggered start of treatments). A simple comparison of this sort may not be sensitive to the effects of the hormone treatment as growth due to treatment may be obscured by normal growth. In other words, during periods when considerable change is normally taking place, it may be necessary to use more refined analyses if supranormal effects are to be detected. Otherwise, measurement is likely to reflect changes which, for the most part, are due to normal development (Figure 2).

Figure 2: Hypothetical effects of growth-stimulating hormone on babies' length

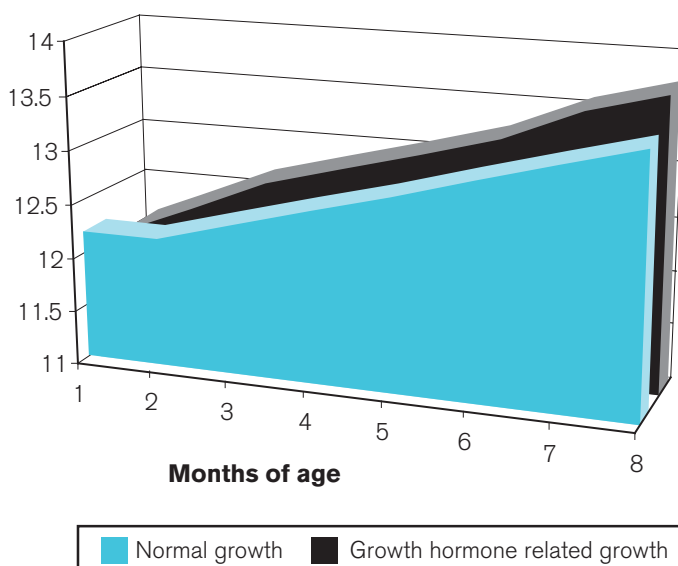


Figure 2 illustrates this hypothetical situation (the figures on the vertical axis are arbitrarily chosen) and shows how a straightforward comparison of growth for the two groups of babies may fail to show the relatively small yet distinct additional growth due to hormone treatment. Transposing this scenario to that of mothers in the twelve months after birth and imagining the growth curve in reverse as representing gradually declining rates of maternal anxiety due to normal adjustment in comparison with adjustment assisted by Home-Start, it seems feasible to speculate that, unless normal changes are partialled out, the small, barely perceptible, changes due to the intervention might well be lost. The next important question, of course, would be how big differences would need to be to create significant long-term benefits for a treated group.

In several further respects, the Birkbeck Home-Start study has some unusual features. Perhaps the most distinctive aspect concerns the strategies adopted to cope with difficulties in recruiting suitable families to make controlled comparisons. As mentioned above, the intention was to compare families that were offered Home-Start as a proactive intervention with those offered Home-Start through the more usual referral routes and those not offered the programme. Difficulties associated with recruiting sufficient and appropriate families into the study and comparison groups led to an attempt to trace families who had been offered Home-Start support but who, for a range of reasons, had not received it. As a result, the sample eventually contained 193 families who had accepted Home-Start support and 196 "control" families. Of the 193 families who had accepted support, 96 actually received two or more visits from a Home-Start volunteer while 97 received one Home-Start visit or less (these were therefore defined as 'non-supported'). Research visits were completed at two and twelve months for 178 of the "control" group families, 92 of the Home-Start 'supported' families and 66 of the 'non-supported' families.

On the first assessment, more differences were found between 'supported' and "control" families than between 'supported' and 'non-supported' families. In comparison with "control" families, 'supported' families were larger, mothers were less likely to be or have been employed, parents were more likely to be in professional or managerial classes and to suffer from mental health problems, had higher educational qualifications and fewer extended family members (e.g. grandparents) living close by. The only difference between the 'supported' and 'non-supported' families was that 'non-supported' families lived in areas rated as more deprived on the Jarman Under-Privileged Area scale. This suggested that the non-supported families might provide a better

comparison. However, it is perhaps important to bear in mind the possibility of systematic differences between 'supported' and 'unsupported' families, for example, with regard to capacity for self-help.

Several widely used measures of maternal wellbeing were employed, including three scales for identification of depressive disorders, a parenting stress index and mothers' perceptions of social support; at twelve months, mothers were also asked whether they had experienced any major life events in the previous twelve months and were interviewed about the nature of their social network. Assessments were also made of the home care environment, feeding practices, use of health services, mother-infant interaction and discipline strategies, infant temperament, health/illness and cognitive development.

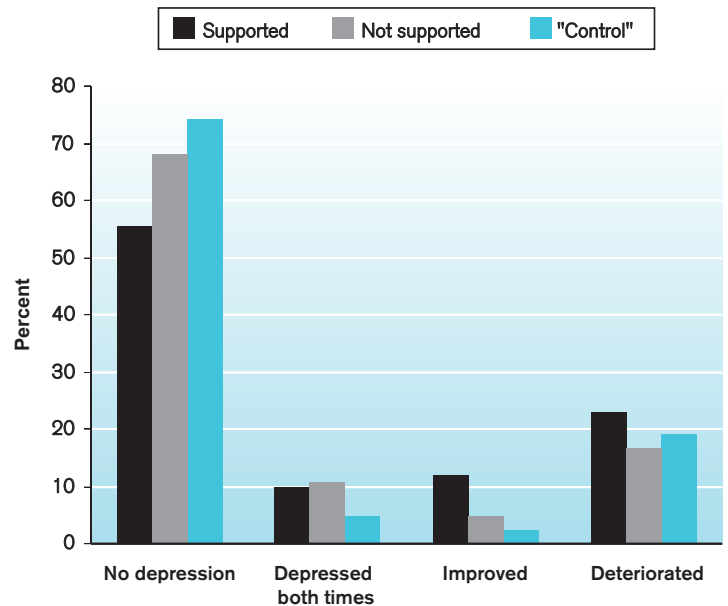
In terms of the specific benefits that Home-Start support might be expected to confer, 'supported' mothers reported more informal support in comparison with either "control" or 'non-supported' families at the time of the second research visit. However, overall, the authors found few benefits for Home-Start supported families and noted that one or two of the small number of differences found appeared to be in a negative direction, for example, 12 month-old children of 'supported' mothers scored lower on Bayley's Mental Development Index than children of 'non-supported' or 'control' group mothers. With such a large array of tests and the likelihood of chance findings, the reliability of these findings remains to be established.

Conclusion

It is difficult to draw any very firm conclusions from either the Belfast or the Birkbeck studies, principally because neither was a randomised controlled trial, both involved families in very diverse situations and neither has yet followed up the families for longer than one year. Also, it has to be noted that the Birkbeck study did not evaluate Home-Start in its usual form and has not yet reported analyses of comparisons between families that received proactive interventions and those that received the more usual Home-Start intervention.

Looking just at the data relating to depression at two and 12 months, it seems evident that the groups were not equivalent (Figure 3). Mothers in the 'supported' group reported higher levels of depression overall, were slightly more likely to improve over the ten months but were otherwise fairly similar to the other mothers.

Figure 3: Differences between groups in Birkbeck sample



Although we would argue that the null findings of the Belfast and Birkbeck studies cannot be taken to be reliable indicators of the lack of effectiveness of Home-Start as an intervention, both studies are of value in pointing to the problems involved in evaluating early intervention programmes. They also both gathered valuable qualitative material which affords useful insights into which families find this kind of intervention helpful and why, and which families might be better served with alternative assistance. However, neither study seems satisfactorily to have addressed the real and very considerable challenges involved in evaluating a programme that offers assistance to such a diverse range of families. Both research teams appear to have started from the premise that all families should be thought of as aiming for the same end goals and that all benefits can be conceptualised as uniform across all families. Given a developmental model in which goals are changing and multifinal, this premise seems questionable. Just as progress for a mother who has suffered from agoraphobia for ten years is unlikely to be indicated by the same outcome as that for a mother who has been a sufferer for two months, so progress for a mother of twins is unlikely to be measured in the same terms as that for a mother with multiple sclerosis. Community samples of vulnerable mothers seem likely to differ considerably from clinical samples selected on the basis of similarity of presenting problem or diagnosis. Calibration of relevant measurement instruments needs to take this diversity into account.

PALS - PRIMARY AGE LEARNING STUDY

The Primary Age Learning Study (Scott et al. 2006) was offered to parents of children in selected Reception and Year One primary school classes in four schools located in one of the poorest parts of London. It consists of 18 school-based sessions (one morning per week) made up of 12 sessions from the Webster-Stratton Incredible Years programme and six sessions designed to teach parents to help their children begin to read. At least one home visit is also offered during the course.

The Incredible Years component, which was originally developed to help parents to manage children with conduct disorders, is based on a social learning (behaviourist) understanding of development, combined with a humanistic approach. It is delivered through a video that models positive parenting skills and ways of coping with children's misbehaviour. Parents watch and discuss the video as a group. The reading readiness component includes the '*Pause Prompt Praise*' approach to reading and teaches parents language and sound exercises to practise and play with their children. The 18-session PALS intervention is a modified version of a previously tested 28-session SPOKES programme found to be effective in similar areas of disadvantage in London (e.g. Scott et al. 2001).

The PALS project also included an element in which parents were videoed in interaction with their children. This observational element, which Scott refers to as a 'gold standard' for intervention trials, involved ten minutes of child-led play, ten minutes on a parent-led construction task and three minutes on tidying up.

Three dimensions of parenting behaviours were rated:

- Parents' 'attachment-promoting style'
- Child-centred and child-directed parental behaviours
- Discipline and positive parenting

At the time of this report, only half of this video material had been analysed due to funding constraints.

Further measures of children's behaviour were based on a semi-structured interview with the parents, completion of the Strengths and Difficulties Questionnaire by parents and teachers, direct observations focused on children's attentional strategies and literacy assessed with the British Abilities Scales single word reading test.

The PALS project was a randomised controlled trial which took place between 2001 and 2004 in four primary schools in the most disadvantaged ward in Southwark, London (Southwark is the third most disadvantaged local authority in

England). In the four schools, 24 classes over three years (672 children in total) participated in the study, with half of the classes in each school randomly allocated to the intervention and half to the no-intervention condition.

At stage one of the study (screening), 665 teachers and 532 parents of Reception and Year One children completed the Strengths and Difficulties Questionnaire and a rating scale based on the diagnosis of oppositional defiant disorder taken from DSM-IV criteria (this yielded data from both parents and teachers on 521 individual children). Summed scores from these were used to identify children most at risk of later social exclusion and low academic achievement (36 per cent or 185 children were identified as above the cut-off point for risk).

At the next stage, in each year of each school, an intervention and control class was randomly selected. All parents were sent letters (details of information in letters are not given) and all parents of children in the intervention classes were offered the SPOKES programme, regardless of their child's rating on the screening questionnaires. Of the 185 children identified as at risk, 87 were in intervention classes and 98 in control classes. All parents of children identified as at risk were offered the programme if they met the inclusion criteria. The appended participant flow chart indicates that 62 of these parents were selected into the intervention, of whom 48 started and 39 completed the programme (according to the flow chart, 87 parents were not selected into the study at this point but as, according to the text, all parents of children scoring above the cut-off point were invited to participate, the '87 not selected' can be assumed to be an error). Of the 336 children scoring below the cut-off point for risk, 165 were in intervention classes and 171 in control classes; the parents of 52 of these children were selected into the programme (41 started and 35 completed the programme); the parents of 62 children were selected as controls (51 started and 45 completed). Therefore, a total of 152 parents provided complete data sets for the study with 74 in the intervention arm (including 39 parents of children identified as at risk of behaviour disorder, i.e. 52.7 per cent) and 79 in the no-intervention arm (including 33 parents of children identified as at risk of behaviour disorder, i.e. 42.0 per cent).

The primary carer from around three-quarters of the families in the study was from a minority ethnic background. There were more boys (52 cf 43 per cent), more lone parents (56 cf 45 per cent) and more parents on household incomes of less than £175 per week (43 cf 34 per cent) in the intervention than the control group. Though age was matched, children in the intervention group had slightly lower

reading scores than those in the control group (7.4 cf 8.1). The characteristics of participants in each group cannot therefore be said to be strictly equivalent though it is not clear how these differences might be expected to influence results.

Few parents attended all 18 sessions: the average number of sessions attended was less than half (7.3) and few significant differences emerged on most comparisons except among parents who had attended five or more sessions. Thirty-one parents attended five or more sessions and from the observational data available for 15 of these, there did appear to be some benefits of programme attendance. They were rated higher on sensitive responses to their children and on child-centred behaviours. No difference was found, however, in respect of child-directive discipline. Analysis of interview data indicated that among all 27 parents who attended five or more sessions, there was less evidence of criticism of children post-intervention while among all 66 parents in the intervention arm there was more evidence of calm discipline. No effects were found in respect of use of praise, suggesting that observational measures may have been more sensitive than interview measures to actual changes.

Few differences were found in relation to outcomes for children: although there was some evidence of differences between intervention and non-intervention groups on some aspects of attention, these differences did not appear to be consistent across interview and observational data.

Effects did not seem to be associated with ethnicity although there was a slight tendency for more white British than African and African Caribbean parents to both take up and attend more sessions (11/15 white British parents in comparison with 17/57 African and African-Caribbean parents attended five or more sessions). The predominant reason given for non-attendance was other commitments.

While a larger sample is necessary to establish the reliability of these findings, it seems that the requirement to spend one morning a week undertaking parenting classes is likely to pose difficulties for most parents, whether or not they are lone parents, have other children, paid work, housework or other commitments. This indicates the importance of including an assessment of the effectiveness of accessibility in evaluations of parenting classes: When this aspect is not included, it can compromise a comprehensive overall evaluation.

SURE START

The Sure Start programme is rather unusual as an intervention programme in that it was deliberately set up so that participants could have as much say as possible in programme content. This means that each Sure Start Local Programme (SSLP) is unique to its particular local area. In other words, although it is conceived as an intervention offered universally to all residents with children aged 0-4 within a specific local area, in practice, it translates into many different types of activity. This is consistent with the overall aims and principles of Sure Start which sets out, on the basis of evidence from best practice in early intervention, "To work with parents-to-be, parents and children, to promote the physical, intellectual and social development of babies and young children – particularly those who are disadvantaged – so that they can flourish at home and when they get to school, and thereby break the cycle of disadvantage for the current generation of young children".

Taking lessons from best practice, Sure Start aims to improve outcomes for children by enriching the social context in which they and their families reside by:

- Improving the quality and, where necessary, quantity of services available to parents and young children
- Improving access to services
- Facilitating early uptake of services.

In this way, Sure Start aims to:

- 1 Improve health and emotional development for young children
- 2 Support parents in their role as parents and in their aspirations towards employment
- 3 Increase the availability of quality childcare for all children.

Although there is local variability in how the programme operates, SSLPs offer a common core of five services which include:

- Home visits
- Support to parents and families
- Services that support good quality play, learning and care
- Primary and community health and social care
- Access to specialist services

Also, theoretically at least, certain fundamental Sure Start principles stipulate that Sure Start services must be:

- Set up to work in close consultation with and for parents and children
- Engage people in the local community
- Offer services for everyone in the geographically targeted areas

- Be flexible and attractive at the point of delivery
- Start very early (antenatally where possible)
- Be respectful and transparent
- Be community driven and professionally coordinated
- Involve multi-agency 'joined-up' work
- Be outcome driven.

Table 2: National roll-out of Sure Start local programmes (SSLPs)

Trailblazer SSLPs	SSLPs		SSLPs2B	
Not included	Round 1 (n=59)	1999 April	Round 5	n=50
in evaluation	Round 2 (n=69)	2000 June	Round 6	
	Round 3 (n=65)	2001 March	Round 7	
	Round 4 (n=67)	2001 September	Round 8	By March 2004
n=60	n=260		n=264(?)	

Different specific outcomes are associated with different stages of development (e.g. encouragement of breast-feeding for newborns and discouragement of smoking for pregnant and lactating mothers). However, all relate ultimately to short-, medium- and long-term improvements in psychosocial adjustment, general health and in capacity for self care for more than one generation of family members across a raft of social contexts. In this way, it can be seen that Sure Start is rooted in an ecological systems concept of development which takes the view that effective and sustained change is unlikely unless interventions permeate as many levels of individual functioning as possible, from close family relationships through to the extended family and the social world beyond the family.

The National Evaluation of Sure Start

The National Evaluation of Sure Start (NESS) was therefore designed to assess the effectiveness of Sure Start not just in terms of its impact on individuals and their families but also in respect of delivery and organisation of services within the targeted areas. It involves five tasks in relation to evaluation:

- 1 Analysis of the local context of SSLPs
- 2 SSLP Implementation
- 3 SSLP Cost-effectiveness
- 4 Support for SSLP self-evaluations
- 5 SSLP Impact

Each of these five aspects, which the NESS team refers to as modules, is characterised by a different mix of quantitative and qualitative methods of investigation and occurs within the context of local programmes that have been rolled out gradually in an increasing number of areas (Table 2). Each area had been identified, using the 'Jarman' index of social disadvantage, as being among the 20 percent most under-privileged areas in England. The first 260 local programmes which were rolled out in four rounds between 1999 and 2001 have formed the basis for the initial stage of evaluation.

The local context analysis

In the initial analysis of the local context of SSLPs, the 260 SSLPs in rounds 1-4 have been compared with the 50 SSLPs-To-Be. Assessments based on archival records and relevant statistical data are used to provide a picture of the local area and its population.

Examples of the kinds information collected and inspected include:

- Demographic information (age, ethnicity, family make-up, head of household, size, etc.)
- Economic characteristics (employment, worklessness, benefit claimance, poverty, etc.)
- Crime rates
- Adult health
- Health and development of children
- Specialist services, e.g. for vulnerable, at risk, disabled, etc.
- Other local services.

Preliminary findings: local context effects

Initial examinations indicated that SSLP and SSLP-to-be areas were not identical: overall, SSLP-to-be areas contained a higher proportion of residents from minority ethnic backgrounds, fewer white and fewer English-speaking residents, with fewer educational qualifications. This applied both to parents with nine-month old children where differences were found on children's ethnicity, language spoken at home, home income and maternal education and to parents with 36-month old children who differed on all these variables and on mother's occupational status. These differences are further discussed in the impact evaluation (NESS 2005, FR013).

SSLP Implementation

Much of the information gathered in the implementation module is of a descriptive nature and focuses on SSLPs in Rounds 1-4. Data has been collected in three ways by:

- Three annual surveys of the 260 SSLPs in Rounds 1-4
- Case studies of 26 SSLPs reflective of a variety of geographical areas and organisational structures
- A series of themed evaluations, e.g.
 - Maternity services
 - Work with fathers
 - Improving parent employability
 - Play and learning activities
 - Buildings in SSLPs
 - Practice guidance.

The focus of interest in this module has spanned many aspects of the establishment of Sure Start local programmes including, for example:

- How programmes are managed and co-ordinated
- Whether access to services has changed
- Uptake of services, consultation, participation
- The quantity of services offered (e.g. number, intensity, length)
- The quality of services offered (e.g. activities, appropriateness)
- Staffing and other resources
- Collaboration and partnership work
- Community involvement.

Evaluation of the implementation of SSLPs has been complicated by new government initiatives, such as the roll-out of Children's Centres and the National Childcare Strategy, which have impinged upon the process of assessing the development of SSLPs¹.

Preliminary findings: implementation

There is not space to document here the full range of findings from this module but some of the more noteworthy observations might include the following:

- It took longer than anticipated to find suitable premises for SSLPs
- Setting up SSLPs took longer than anticipated
- It takes approximately four years for SSLPs to 'bed down'
- There is considerable variation across SSLPs in services offered
- Most SSLPs have one or two dedicated buildings with additional satellite premises

- One of these buildings usually provides day care
- SSLPs have increased the availability of quality childcare
- At least half of SSLP buildings are shared with other agencies
- Although there is a standardised system of security across SSLP buildings, there is otherwise no one distinctive SSLP style
- Around two-thirds of SSLPs offer maternity services
- There have generally been difficulties in getting fathers involved
- Successful engagement of fathers is more likely if SSLPs are specifically geared towards fathers
- SSLPs and mainstream maternity services took time to begin to work together
- Proliferating SSLPs presented resource problems for voluntary and statutory organisations
- National skills shortage (e.g. of health visitors and midwives) complicated SSLP development
- Complementary services have been successfully established
- SSLPs have enriched service delivery and practice
- SSLPs are being used as gateway to parental work/training
- SSLPs are also being used as gateways to other services.

There have been distinct challenges for SSLPs in initiating partnership work for a number of reasons. Chief among these is the fact that multi-agency working is inevitably more complex and labour-intensive than solo work and involves a new approach to service delivery for many providers. Concerns about the sustainability of Sure Start initiatives and a sense that SSLPs may have 'hijacked' existing agencies and 'poached' skilled workers who were already in short supply have not been uncommon: "I think there's some fantastic stuff coming out of Sure Start. However, it gets under my skin that all these years we've been doing this stuff and now Sure Start gets all the credit" (*Family Support Coordinator, Large national voluntary agency: NESS 2005, FR010*).

The notion that Sure Start purported to offer a brand new sparkly approach that was to put all that had gone before in the shade may well have created some resistance to the building up of effective partnerships in the initial stages. Even more seriously, it also seems possible that the assumption that Sure Start was to be so new and different may have pre-empted consideration of a full evaluation of the way that Sure Start was to add to or modify services, particularly those provided by voluntary organisations that were already at work and competing for funding in Sure Start areas. There appears not to have been any systematic or detailed audit of pre-existing provision or of the effect upon this of Sure Start,

¹ Within the Ten Year Childcare Strategy, from April 2007, funding mechanisms will change significantly so that local authorities will receive Sure Start funding to distribute locally through Children's Centres.

except from the perspective of the SSLPs themselves. Costly though such an audit might be, for a thorough assessment to be made of the processes involved in establishing Sure Start within the context of existing provision and to gain a complete picture of the place of Sure Start programmes within this wider context, more attention might usefully have been given here.

Perhaps consequent upon this or, if not, certainly in a similar vein, evaluation of services uptake and use appears to have received relatively little attention within the Sure Start evaluation. On this aspect, evaluation of Home-Start appears to have involved a more systematic attempt to assess the extent to which participation may have influenced service uptake. Utilising an instrument especially designed to assess the nature of services and their use (the Client Service Receipt Inventory), which takes approximately 15 minutes to administer, Sleed et al. (2005) gathered information about which services had been used, how much and how this had changed over the course of the Home Start programme. NESS respondents appear only to have been asked to state whether they have ever sought advice in relation to a list of topics or issues. This gives some indication of the range of concerns of parents in SSLPs, but cannot capture much information about how the Sure Start programme may have changed the overall pattern of services on offer in SSLP areas.

SSLP Cost-effectiveness

There are a number of reasons for choosing not to discuss this aspect at this point. First, although data is being collected on cost-effectiveness, it is still rather too early in the life of SSLPs and their clientele to draw any confident conclusions about Sure Start overall. Second, the overall long-term aim of Sure Start, in reduction of the risk of social exclusion and poverty, is a complex ambition which will perhaps be better examined by the longitudinal element in the evaluation. Again, it is too early in the life of Sure Start for these reports to be available.

Meantime, big differences have been observed among SSLPs in costs per child (average around £1,000; range from £300-4,000: NESS 2005, FR010). There has also been some concern from partners both about whether the relatively high level of Sure Start funding is well used and about the amount of money Sure Start projects have in comparison with other projects or areas (NESS 2005, FR010). On the bright side, there is evidence of creativity in use of financial resources, for example, some SSLP managers have managed to secure additional sources of funding or to save money in developing projects, often

through making arrangements with partners to share costs and resources such as buildings. Such arrangements, in the long term, seem more likely to secure sustainability but must clearly add to the complexity of the task for the NESS team in carrying out an effective cost-benefit analysis.

Support for SSLP self-evaluations

Occasional remarks in the NESS reports suggest that the task of supporting SSLP self-evaluations has not been an easy one. The Sure Start website has offered a forum for information exchange, and the NESS team have given additional training and advice, for example, in the form of recommendations about the use of standardised instruments across SSLPs. Reading between the lines of some of the themed evaluations (e.g. Myers et al. 2005; NESS 2005, FR012), it would appear that the overall quality of self-evaluations has been disappointing, and that information emerging from these has so far been less consistent or reliable than may have been anticipated.

Impact of SSLPs

Evaluation of the impact of SSLPs aims to discover short-, medium- and long-term outcomes for children and their families of two kinds: specific effects on outcomes and what works best for whom.

Impact evaluation was to proceed in two phases:

- 1 A preliminary cross-sectional comparison of 150 SSLPs and 50 SSLPs-to-be
- 2 A longitudinal study of 8,000 families with children aged two and four from 100 of the 150 SSLPs.

The longitudinal study was to track a subsample of SSLP families and compare outcomes for these people along relevant variables from other large population surveys, such as the Millennium Cohort Study, and the General Household Surveys. So far, preliminary findings have only been reported from a cross-sectional study of parents of 9 and 36 month-old children three years after the initial roll-out of Sure Start. Recruitment into the cross-sectional study was principally via random selection of parents with 0-4 year olds on Child Benefit registers relevant to the 150 Round 1-4 SSLPs and 50 SSLPs-to-be. This process yielded 16,502 children in SSLP areas and 2,610 in SSLP-to-be areas (Table 3), an 84 per cent response rate for parents of nine month-olds and a 73 per cent rate for those of 36 month-olds.

Table 3: Impact study (NESS November 2005; FR013)

SSLP		SSLP-2B	
9 month	36 month	9 month	36 month
3,927	12,575	1,509	1,101
16,502		2,610	

Assessments, based on home visits and follow-up telephone interviews, were carried out to explore differences between children and their families in the 150 SSLPs and the 50 SSLPs-to-be. These employed standardised measures or derivatives of these to monitor salient dimensions:

Child development

Cognitive development

(using 4 subscales of the British Ability Scales)

- Block building
- Picture similarities
- Verbal comprehension
- Picture naming

These indicate verbal and non-verbal ability

Psychosocial adjustment

(using dimensions of the SDQ)

- Social competence
 - Pro-social behaviour
 - independence
- Behaviour problems
 - conduct
 - hyperactivity
 - emotional control
 - overall difficulties

Quality of care

- **Home as care environment**
(widely used HOME instrument)
- **Home as learning environment**
(newer Home Learning Environment measure)
- **Home chaos**
(questions devised to assess home organisation)
- **Harsh discipline**
(items from Parent-Child Conflict Tactics Scales)
- **Parent-child conflict**
(items from Student-Teacher Relationship Scale)
- **Father involvement**
(items from Millennium Cohort Study)

- **Maternal self-esteem**
(a 6-item subscale derived from standard measures)
- **Maternal malaise**
(a 9-item subscale)

Since additional fairly extensive demographic information pertaining to parents and children in the family also needed to be collected on these home visits, efforts were made to ensure that the question load was kept as low as possible. Thus, relatively few complete standardised measures were used. Instead, subscales or items derived from standardised instruments were selected. This approach can sometimes compromise reliability.

A perhaps more serious threat to reliability, where large numbers of measures are taken and a lot of statistical tests are carried out, is the risk of type one (false positive) errors (since, by chance, significant effects will occasionally occur and increasing the number of tests increases the number of chance findings). To lessen the likelihood of this threat, the number of analyses required was reduced using factor analytic techniques. In this way, although results for the home learning environment variable were treated separately, outcomes on all the other child and family measures were reduced to four dimensions (Table 4).

Table 4: Dimensions of child and family outcomes

Supportive parenting	Child social competence
responsivity	Pro-social behaviour
acceptance	independence
Negative parenting	Child emotional/behavioural difficulties
parent-child conflict	conduct problems
harsh discipline	hyperactivity
home chaos	emotional dysregulation
(inverse) parent-child closeness	overall difficulties

Preliminary findings: impact

Several analyses were run on the data that resulted. First, only complete data sets were analysed. Second, cases with missing data were included by using information about overall performance to impute missing scores (which ranged from 10-41 per cent of scores throughout all variables). Next, to account for the differences between SSLPs and SSLPs-to-be that were mentioned in connection with the local

context analysis (above), child, family, background and area factors were added to the model. Thus, six sets of results were inspected, three pertaining to families with 9 month-olds and three to families with 36 month-olds (Table 5).

Table 5: Data sets (NESS November 2005, FR013)

	Complete data sets		Imputed data sets
	Unadjusted	Adjusted	Adjusted
9 month	*	**	**
36 month	*	**	**

** results considered most important

At this point, a further decision was taken to guard against the possibility of type one errors: the only findings of difference between SSLP and SSLP-to-be that would be accepted as valid would be those arising from analyses of both the complete and imputed sets of adjusted data for each age group.

On this basis, the main findings were that:

- In SSLP homes with 9 month-olds, there was less household chaos
- In SSLP families overall, scores were lower on home chaos
- SSLP Mothers of 36 month-olds were more accepting towards their children
- SSLP Mothers of 36 month-olds, with the exception of teenage mothers, evinced less negative parenting
- SSLP 36 month-old children of non-teenage mothers were rated
 - higher on social competence
 - lower on behaviour problems
- SSLP 36 month-old children of teenage mothers were rated
 - higher on behaviour problems
 - lower on social competence
 - lower on verbal ability
- SSLP 36 month-old children in workless homes and in lone parent families were also rated lower on verbal ability.

Overall, it was concluded that there was little evidence of any positive impact of Sure Start. Very few main effects had been found and, where they were, effect sizes were small. Looking at specific sub-populations, it appeared that, particularly among mothers of 36 month-olds, Sure Start may have produced only modest benefits for the less disadvantaged families. For the more disadvantaged families, such as lone parents and teenage mothers, that is, precisely those families

whose futures the programme was set up to change, Sure Start appeared not only to be failing to help but to be having a negative impact. Clearly, this was a cause for concern.

Further analyses suggested that these patterns may be associated with variations in the implementation (progress and processes) and philosophy of individual SSLPs. Eighteen dimensions along which programmes varied were identified (Table 6). These dimensions were not discrete, or mutually exclusive, categories but were considered capable of collectively distinguishing between effective and non-effective SSLPs on child and parenting outcomes.

Table 6: Dimensions of programme implementation

[A] Progress with implementation (7 dimensions)	[B] Processes underlying implementation (7 dimensions)	[C] Holistic aspects (4 dimensions)
service quantity service delivery identification of users, reach reach strategies service innovation service flexibility	partnership composition partnership functioning multi-agency working leadership access to services evaluation use staff turnover	vision communications empowerment ethos

Analyses using programme ratings along these dimensions indicated the following patterns:

For parents of 9-month-olds		
Empowerment by SSLPs	is associated with	More maternal acceptance
For parents of 36-month-olds		
Empowerment by SSLPs	is associated with	A more stimulating home learning environment
Better identification of users	is associated with	Greater non-verbal ability
Stronger programme ethos	is associated with	More maternal acceptance
Higher % health-related staff	is associated with	Maternal acceptance
Improvement in child services	is associated with	Maternal acceptance
Inherited parent-focused services	are associated with	Less negative parenting

SSLPs led by Health agencies generally appeared to be more effective than other statutory or voluntary agencies. For example, they were associated with greater involvement among fathers of nine month-olds and with lower numbers of accidents among 36 month-olds. Also, mothers of both nine and 36 month-olds living in SSLPs led by health agencies tended to rate their area as more satisfactory for bringing children up.

DISCUSSION

Problems for evaluation: what are we measuring?

It is difficult to know what sense to make of these early findings. They certainly raise questions about the processes involved in the National Evaluation of Sure Start and its scope. For various reasons, it seems likely that interpretation of findings will continue to be problematic. Taking the findings above as an example, it is not possible to tell whether services in these areas were, before the advent of Sure Start, more satisfactory or whether SSLPs have contributed². There must, surely, also be some question about the notion that Sure Start schemes should have the capacity to change the overall nature of under-privileged areas. However brilliant a Sure Start initiative may be, can we seriously believe that it alone could make local areas rated as the worst places in the UK to live in change in a short space of time into good places in which to bring up children? It is also important to be aware of the well-established effect shown in the literature on effects of empowerment, that increase in expressed negativity may be associated with the positive effects of empowerment in the early stages of effective intervention. In the case of Sure Start, this might have meant that previously depressed or resigned mothers, are empowered, may have felt supported in speaking out and demanding better services.

NESS: The need for caution in drawing potentially premature conclusions from early findings

Although ideally an RCT design may have been preferable, in its absence and as a compromise solution, the NESS design, in many ways, appears to have been carefully crafted and many checks have been put in to avoid over-estimation of

programme effects. Some may argue that this may have put it at risk of under-estimating effects (though others may also point to other aspects, such as the allocation of programmes to areas, which may have resulted in an over-estimation). However, it does seem curious that, given the care that was taken to avoid type one (false positive) errors and the fact that one of the strategies employed to achieve this end was the factor analytic reduction of single variables to four dimensions, few of the reported outcomes appear to relate to the four factors. Rather, they relate to single variables such as home chaos, maternal acceptance, or verbal ability.

It also seems curious that the patterns of negative outcomes for sub-groups closely resemble the outcomes that might have been expected had local context effects not been adequately controlled for. The local context analysis indicated that in SSLPs-to-be there were more BME families, fewer children with English as a first language and families with lower educational qualifications than in SSLP families. These differences can be expected to result in poorer verbal ability, if verbal ability is indexed by use of English, and, possibly, a tendency towards more authoritarian parenting styles, since these have been found to be associated with larger families and with families from communities that are less urbanised (Kagitcibasi 1996). Is there, perhaps, a possibility that when split file analyses were carried out, controls for context effects may have been omitted and that these are the differences that have been picked up by the sub-population analyses? While this seems unlikely, it is a consideration that is worth investigating further, given the impact that NESS findings may have on political decisions about the provision of children's and families' services.

From ONS statistics (Barrett 2004), it is also evident that if the population of teenage mothers is set as mothers under 18 or 20 and if mothers of 17 and above are treated as identical with those under 17, there is a possibility that younger marrying Asian women may be over-represented in this group. These considerations too suggest that it may be valuable to explore the nature and source of differences if their significance is to be fully understood.

Caution in drawing conclusions about these preliminary findings is also recommended on other grounds. The NESS design had to meet at least four major challenges and each of these can make for difficulties in interpreting its findings:

- First, the requirement to offer Sure Start to all 20 per cent of areas rated highest on indices of social disadvantage over a relatively short space of time and the time that it has taken for programmes to 'bed down' have created problems for evaluation.
- Second, the need for SSLPs to be user-driven and to

² It is also important to note in this connection that in some areas defined as high on the index of social disadvantage, existing initiatives such as urban regeneration schemes would already have been operating. Therefore, in some local authorities, decisions may have been taken to offer the SSLP to a neighbouring area, still appropriately high on social disadvantage but not the most deprived ward in the borough. Analyses do not appear to have allowed for the fact that there was no consistent strategy for allocating Sure Start programmes in these situations and it is therefore possible that this may have compromised the comparability of SSLPs and SSLPs-to-be.

facilitate enormous programme variability has created a situation where moving targets and multiple goals make it very difficult to adopt a simple design with just a few standardised measures.

- Third, the onset of other major government initiatives, such as the roll-out of Children's Centres, has added a moving playing field to the moving targets and multiplying goals. Longitudinally, this means that effects of Sure Start may become obscured by the addition of other initiatives.
- Fourth, in spite of the enormous sums of money involved in initiating and evaluating Sure Start, various arguments appear to have militated against setting it up as a randomised controlled trial. This has meant that, although in theory it still might have been possible to match individual families in a matched pairs design and to have controlled for all relevant contextual or background factors (such as parents' income, education, family size, health, etc.), in practice, with a project of this scale, this could have proved even more expensive and difficult to organise. Instead, the evaluation has adopted an 'intention to treat' design which has also posed problems.

circumstances regardless of whether or not treatment is offered. For example, how appropriate is it in situations where cases in need of treatment have not been identified, where they are present only in a small minority of the total population offered treatment, or where non-compliers substantially outnumber compliers? Extending a medical paradigm too far beyond the population or circumstance for which it is designed risks over-extension. Over-extended, it may back-fire. Typically, in the clinical setting, the majority of patients comply with recommended courses of medicine and a substantial proportion of those most seriously in need is therefore included in the treated sample. Only a minority of patients do not receive the treatment at all. This situation cannot be assumed to obtain where early intervention programmes are offered universally for a number of reasons, for example, because (a) it is known that the families most in need are those most likely to refuse to participate, (b) not all families would necessarily be expected to benefit from taking part since some may already have the skills offered or may need alternative programmes more suited to their individual requirements, (c) typically, few potential recipients participate in the full programme and (d) often, no clearly identifiable 'course' or programme, with measurable outcomes, is on offer.

How widely applicable are 'intention to treat' designs?

No published NESS report so far has documented the proportion of SSLP area respondents that has actually participated in Sure Start programmes. The reason for this is that an 'intention to treat' approach is central to the NESS design. Intention to treat designs treat all those assessed as in need of and offered treatment as if they were in receipt of treatment. This approach is well-established within clinical intervention trials and is thought to give a more accurate estimate of programme effectiveness because it provides information about all potential patients, rather than only those who are treatment-compliant (systematic differences have been found between treatment-compliant and treatment non-compliant patients). For example, it avoids over-representation of the subgroup of patients who are so highly committed to treatment that outcomes for them begin to approximate to placebo effects. This both blurs the distinction between treatment effect and other effects and over-inflates the effect of treatment. Therefore, the intention to treat approach is thought to provide a more conservative estimate of programme effectiveness by reducing the risk of type one (false positive) errors and heightening the risk of type two (false negative) errors. It is also argued that any effect still found after use of this procedure must be real.

However, it does seem important to ask whether the 'intention to treat' design is appropriate either in all circumstances where treatment is offered or in all

Using an intention to treat design in situations where no uniform 'condition' (other than the human condition) has been identified and where no specifically targeted and so measurable 'treatment' is available seems destined to produce at best rather haphazard if not potentially rather bizarre results. It could hypothetically be the metaphorical equivalent of offering walking sticks to all contestants in all races at a school sports day and then assuming that everyone has used them whether or not they have. In reality, under such circumstances, the speed with which the walking stick is rejected might prove to be more strongly associated with race-winning than its retention. In such circumstances, where there is no identified condition to treat, use of an intention to treat paradigm may not only fail to avoid the risk of over-estimating effects but may provide a measure of unwanted or irrelevant effects. This seems particularly likely where uptake of a programme is very low and where potential recipients have very diverse needs.

In the case of analyses of the PALS data, an alternative approach was adopted which combined an intention to treat analysis with a subsidiary analysis that took account of the number of sessions actually attended. This would appear to imply a tacit acknowledgement, at least, of the limitations of the intention to treat approach where programme uptake is relatively low. The intention to treat approach seems justified in the case of PALS because of the intention to influence the

incidence of antisocial behaviour. Nevertheless, there would still appear to be questions around its use where behaviour is to be indirectly 'treated' or changed by altering aspects of the care environment or of other behaviours thought to be related. Such indirect treatment approaches seem likely to be more difficult to assess than interventions that involve clearly identifiable remedies applied to clearly identified conditions.

Where there is no clearly identified condition and no clearly identified 'treatment', or where there are many conditions and many different 'treatments', the use of an intention to treat design seems less appropriate. In such situations, it may be more appropriate to carry out a simple assessment of outcomes in relation to 'dosage' (number and nature of sessions) and condition at outset. Given the time and resource constraints within which the NESS evaluation has operated, this perhaps has not presented as a viable option in the short term. However, where early intervention programmes are offered to very diverse families in very different circumstances, this more fine-grained analysis of outcomes, though more labour-intensive, may well prove to be extremely fruitful in the long term.

Are we asking the right questions?

Numerous evaluations and meta-analyses of evaluations have been carried out on the assumption that the appropriate questions to ask about intervention programmes are (a) whether they work and, if so, (b) which programmes work best. Premised on the understanding that RCTs are the most appropriate format for evaluation, they elicit information about how groups of participants perform on certain outcomes which are assessed, usually, with the use of standardised measures. Almost inevitably, this kind of evaluation tends repeatedly to demonstrate that most interventions work to some extent for most parents. They provide very little information about which elements of programmes have worked best for which parents, which have worked least well, or what may be needed to help parents who are in difficulty but who, for various reasons, do not manage to engage with or benefit from interventions³. Yet this information is often what practitioners are seeking and what they need to make sure that interventions are both maximally effective and cost-effective.

For this reason, we would argue that some key assumptions on which evaluations have customarily been premised may need to be reviewed. For example, the assumption that assessment can be conducted as though the development of

confident and effective parenting skills is analogous to the treatment of a medical condition or an identifiable disorder may need to be questioned. This assumption brings with it an expectation that the effectiveness of interventions can be assessed using tools designed to capture specific outcomes (e.g. increased self-esteem, lower anxiety or depression scores, better behaviour, etc.) and that these effects can be expected to occur within particular, usually rather short, time scales. This kind of approach may be appropriate in some circumstances, for example, where entry into programmes has been due to parents' concerns about children's conduct or where all parents have been selected on the basis of low scores on mood rating scales, for example. In this situation, tangible targets for behaviour change can be set. However, it is unlikely to be similarly relevant where interventions have set out to facilitate less tangible changes or to produce effects that are not expected to be uniform across all participating parents. In these situations, RCT designs may need to be complemented by other means of assessment.

Where interventions are aimed at more diverse populations, the questions that need to be addressed are rarely simply whether a programme works or even how well it works. Programme developers will, usually, routinely have had numerous evaluations that have established the answers to these questions more or less convincingly. These evaluations will not have addressed the more specific and pressing questions concerned with ascertaining how programmes work with families that have complex and multiple problems, namely, questions about which programmes or elements of programmes work best for which kinds of family, in which circumstances and with which kinds of problem.

The need for evaluations more directly tailored to capture effects of interventions on individual families

To answer these more complex questions, in addition to RCTs, evaluations will need to be more directly tailored to capture the situation and needs of individual families and more sensitive to variability in speed of change. To do this, they will need to have greater cognisance of the nature of development and of the multiplicity of routes to progress. In doing this, evaluations are unlikely to consist only of easy-to-administer quantitative measures or to be premised on the assumption that all programme participants start from the same baseline and aim to reach the same goals. The evaluations required will be more likely therefore to utilise a mix of qualitative and quantitative methods, with a strong emphasis on comprehensive delineation of individual targets and needs, accompanied by identification of the programme or programme elements most suitable for particular families. To some extent, both Home-Start evaluations have begun to

³ Additional studies, such as that taken by Barnes et al. in relation to non-engagement with Home-Start (Barnes et al., draft paper), demonstrate the usefulness of complementary data collection.

shed some light on how these questions may be addressed, by focusing on identifying the elements of programmes that do not work for particular families. It may well be that within SSLP self-evaluations and longer term follow-ups, more information of this nature will also be revealed. However, if the emphasis continues to be on 'quick fixes' rather than on the complex support needs of families in widely differing circumstances, answers to these more difficult questions may take a long time to find.

CONCLUSIONS

- (1) **The importance of building sound evaluation into designs:** Perhaps the chief message to be taken from this 'evaluation of evaluations' is the importance of building into social programmes a capacity for sound evaluation. Ideally, the closer evaluations can approximate to randomised control trials, the more likely they are to control for all potentially confounding variables and a clear picture will then emerge of effects due to the programme itself.
- (2) **The need to think outside the 'box' of RCTs:** Even where clean, tidy RCTs are possible, and this is unlikely to be the case in the majority of interventions where sample sizes and funding constraints will make quasi-experimental designs seem preferable, it will often be important to recognise the value of collecting a variety of data from multiple sources. Programmes are likely to be offered to parents with differing developmental trajectories and very variable and complex needs. Understanding the nature of the interaction between programmes and participants may be better captured by supplementing quantitative methods with qualitative methodologies more suited to understanding the mechanism of unique effects.
- (3) **The value of increased knowledge about unique or idiosyncratic effects:** There may be a sense in which RCTs and meta-analyses 'throw the baby out with the bath water'. They provide insights into the effect of programmes on homogenised populations, and about central tendencies as opposed to tail-end effects. Where understanding is sought of more heterogeneous aspects, they are necessarily silent because they deliberately partial these variations out. Yet an understanding of how programmes work, or do not work, with Mr. and Mrs. Extreme and Complex is likely to be equally, if not more, valuable than knowing whether they suit Mr. and Mrs. Average. If help is to reach the most troubled families, knowledge about heterogeneous effects will be needed: this is likely to require systematic collation of information
- about how programmes had to be adapted for particular parents. By identifying commonalities and differences within information about processes of programme delivery, the important insights needed by practitioners into what works for whom can be acquired.
- (4) **The danger of premature foreclosure on the process of evaluation:** It hardly needs saying that for many troubled families, change takes time. Building relationships with service providers requires continuity of provision over long periods. It follows that evidence of change is unlikely to emerge quickly and is likely to take many forms. Direct effects as well as ' sleeper' effects may not be visible for years. This is a well-established phenomenon. Even if effects are found soon after a programme has been run, this is no proof that programmes will retain these effects in the long term. Early findings need to be treated with great caution and not seen as being capable of forming the basis for dramatic policy changes.

References

Barnes, J., MacPherson, K. and Senior, R. (in preparation) Factors influencing the acceptance of volunteer home visiting support offered to families with new babies.

Barnes, J., Senior, R. and MacPherson, K. (in preparation) Right from the Start: Evaluation of Home-Start with mothers of newborn infants. Part 1: Quantitative results. Final Report to Joseph Rowntree Foundation.

Barrett, H. (2004) *UK Family trends 1994-2004*. London: National Family and Parenting Institute.

Daily Mail (September 14, 2005) *A less than sure start*. Article by Melanie Phillips.

Glass, N (1999) Sure Start: The development of an early intervention programme for young children in the United Kingdom. *Children and Society*, **13**, 257-264.

Guardian (September 13, 2005) *Doubts over value of £3bn Sure Start*. Article by Lucy Ward.

Guardian (September 13, 2005) *We must hold our nerve and support deprived children*. Article by Polly Toynbee.

Kagitcibasi, C. (1996) *Family and human development across cultures: A view from the other side*. Hillsdale, NJ: Lawrence Erlbaum.

McAuley, C. (1999) *The Family Support Outcomes Study*. Northern Health and Social Services Board. Ballymena, Northern Ireland.

McAuley, C., Knapp, M., Beecham, J., McCurry, N. and Slead, M. (2004) *Young families under stress: Outcomes and costs of Home-Start support*. York: Joseph Rowntree Trust.

Scott, S., O'Connor, T. and Futh, A. (2006) *What makes parenting programmes work in disadvantaged areas? The PALS trial*. York: Joseph Rowntree Foundation; London: Institute of Psychiatry.

Scott, S., Spender, Q., Doolan, M. and Aspland, H. (2001) Multicentre controlled trial of parenting group for childhood antisocial behaviour in clinical practice. *British Medical Journal*, **323**, 1-7.

Scott, S., Sylva, K., Doolan, M., Jacobs, B., Price, J., Crook, C. and Landau, S. (in press) Randomised controlled trial of parenting groups targeting multiple risk factors for child antisocial behaviour: the SPOKES project.

Slead, M., Beecham, J., Knapp, M., McAuley, C. and McCurry, N. (2005) Assessing services, supports and costs for young families under stress. *Child: Care, Health and Development*, **32(1)**, 101-110.

Sunday Times (September 18, 2005) *Pity the poor children left to Blair's care*. Article by Minette Marrin.

Bibliography: Sure Start reports

Tunstill, J., Allnock, D., Meadows, P. and McLeod, A. (2002, June) *Early experiences of implementing Sure Start*. Nottingham: DfES. [NESS/SF/001]

Ball, M. (2002, June) *Getting Sure Start started*. Nottingham: DfES. [NESS/FR/002]

Barnes, J., Broomfield, K., Frost, M., Harper, G., McLeod, A., Knowles, J. and Leyland, A. *Characteristics of Sure Start local programmes: Round 1 to 4*. Nottingham: DfES. [NESS/SF/003]

Lloyd, N., O'Brien, M. and Lewis, C. (2003, August) *Fathers in Sure Start local programmes*. Nottingham: DfES. NESS/SF/004

Full report: National Evaluation Report 04. [NESS/FR/004]

NESS (2004, June) *Characteristics of Sure Start local programmes 2001/2*. NESS/FR/005.

NESS (2004, June) *Improving the employability of parents in Sure Start local programmes*. National Evaluation Report 06. Nottingham: DfES. [NESS/FR/2004/006]

NESS (2004, June) *Towards understanding of Sure Start local programmes. Summary of findings from the national evaluation*. Nottingham: DfES. NESS/SF/007

NESS Implementation Team (2005, January) *Implementing Sure Start local programmes: An in-depth study, Part One*. National Evaluation Report 07. Nottingham: DfES. [NESS/FR/2005/007]

NESS Implementation Team (2005, January) *Implementing Sure Start local programmes: An in-depth study, Part Two – A close-up on services*. National Evaluation Report 07-2. Nottingham: DfES. [NESS/FR/2005/007-2]

Barnes, J., Desousa, C., Frost, M., Harper, G., Laban, D. and the NESS team (2005, July) *Changes in the characteristics of Sure Start Local Programme areas – 2000/2001 to 2002/2003: Summary*. Nottingham: DfES. [NESS/2005/SF/008]

Full report: NESS/2005/FR/008 *Changes in the characteristics of Sure Start local programme areas in Rounds 1 to 4 between 2000/01 and 2002/03*

Myers, P., Barnes, J. and Kapoor, S. (2005, July) *Speech and language services in Sure Start Local Programmes: Findings from local evaluations*. NESS: Institute for the Study of Children, Families and Social Issues, Birkbeck.

Anning, A., Chesworth, E. and Spurling, L. (2005, November) *The quality of early learning, play and childcare services in Sure Start Local Programmes*. National Evaluation Report 09. Nottingham: DfES. [NESS/2005/FR/009]

NESS (2005, November) *Maternity services provision in Sure Start local programmes*. National Evaluation Report 10. Nottingham: DfES. [NESS/2005/FR/010]

NESS (2005, November) *Buildings in Sure Start local programmes*. National Evaluation Report 11. Nottingham: DfES. [NESS/2005/FR/011]

NESS (2005, November) *Maternity services provision in Sure Start local programmes*. National Evaluation Report 12. Nottingham: DfES. [NESS/2005/FR/012]

NESS (2005, November) *Early impacts of Sure Start local programmes on children and families: Report of the cross-sectional study of 9- and 36-month old children and their families*. National Evaluation Report 13. Nottingham: DfES. [NESS/2005/FR/013]

NESS (2005, November) *Variation in Sure Start Local Programmes' effectiveness: Early preliminary findings: Report of the NESS Programme Variability Study*. National Evaluation Report 14. Nottingham: DfES. [NESS/2005/FR/014]

Other Sure Start documents

Sure Start Children's Centres: Practice Guidance.

National evaluation of Sure Start – Summary.

(1999, November) *Sure Start Evaluation Development Project: Report to the Sure Start Unit*.

(2000, June) National evaluation of the Sure Start programme in England. Specification of requirements.

(2001, March) *Parents' Views. The impact of Sure Start – one year on*.

Evaluating Evaluations

Evaluating recent evaluations of Sure Start, Home-Start and Primary Age Learning Study

Helen Barrett

Recent evaluations of two UK early intervention programmes, Sure Start and Home-Start, have been the subject of considerable attention. The programmes have aimed to boost children's life chances by tackling social exclusion and child poverty. This review aims to examine the recent reports, and that of a third programme, the Primary Age Learning Study, and to make an objective as possible an assessment of the significance of their findings.

Evaluating Evaluations is the review written for the seminar: *Commission in Haste, Repent at Leisure? Evaluations of Family Preventative Services and the Implications for the Development of Policy*. It includes discussion of the methodology of evaluations and the nature of the findings, their validity and implications for the purposes of policy influence. It offers a highly informative analysis of family service assessment, and will be of great interest to policy makers, academics and professionals for whom evaluation is a critical factor in developing a new and often contentious service field.