Predicting impact in an Early Years intervention: the design of a tool using qualitative and quantitative approaches

Angela Anning[1], Mog Ball[1], Jay Belsky[1] and Edward Melhuish[1]

[1] National Evaluation of Sure Start team, based at the Institute for the Study of Children, Families and Social Issues at Birkbeck, University of London, UK

Abstract

This article focuses on the design and application of an instrument, the Programme Variability Rating Scale (PVRS), to measure the effectiveness of a complex social intervention in the UK. Sure Start aimed to improve outcomes for children aged under four years living in disadvantaged areas on a wide range of health, educational and social indicators. The PVRS was devised for use in the National Evaluation of Sure Start (NESS) to measure programme proficiency. It consisted of 18 dimensions (eg. parental empowerment, user identification, flexibility of service delivery), each with seven levels of proficiency. It was applied to 150 Sure Start local programmes involved in a longitudinal study of the impact of the intervention on a range of child and parental outcomes. Ratings of more or less proficient processes were related, using discriminant-function analysis, with the impact outcomes from the cross-sectional study of almost 20,000 children. The 18 dimensions of proficiency of the PVRS made a significant differentiation between the most and least effective programmes.

Key words

predicting impact; Early Years intervention; qualitative and quantitative approaches; the Programme Variability Rating Scale

Background

Social researchers have long wrestled with the problem of linking the processes and activities of large-scale social interventions with their intended outcomes. Examples within children's services from the US are evaluations of: the childcare-based Abecedarian Project (Ramey *et al*, 2000); the early education programme Head Start (Love *et al*, 2002); the home-based Prenatal Early Intervention Project (Olds *et al*, 1999); and the behaviour modification programme Incredible Years (Webster-Stratton, 1993). An example from Australia is the evaluation of the family support Triple-P Positive Parenting Program (Sanders, 2003).

Where programmes are delivered with well-defined procedures, protocols and curricula it is easier to compare the way in which they are implemented. For example, the Prenatal Early Intervention Program, which involved nurses conducting home visits to families with new babies over a set period of time, specified clear procedures for the home visits. The Incredible Years model required those delivering treatments to use a manual with a prescribed curriculum. The task of the evaluator is to assess the efficacy of the programme – that is to determine if it is being implemented in the way prescribed – and then to set comparisons of impact against variations in programme delivery. But in more complex interventions, such as Head Start,

where a range of targets were prescribed and treatments were delivered in diverse ways, the variables were more difficult to manage when trying to compare programmes. The evaluations of Early Head Start (for 0–3-year-olds) and Head Start (for 3-year-olds upwards) used the programmes' progression through Head Start Program Performance Standards, which emphasised the centrality of children's development and stressed programme quality through adherence to the standards, as one way of comparing programmes. But evaluators of Early Head Start also divided programmes into early, late and incomplete implementers (based on the principle that length of treatment would predict impact) and centre-based, home-based or mixed approach (based on the principle that different configurations of services would predict different impacts) (Love *et al*, 2002).

The fundamental methodological challenges of measuring outcomes and determining their relationships with processes in the evaluation of interventions reflect debates about the efficacy of scientific/positivist approaches to research and evaluation (where there are perceived objective truths) as opposed to constructivist approaches (where there are as many truths as there are agents). Recent approaches have attempted to synthesise the two extremes and to combine the analysis of quantitative and qualitative data. Examples are realistic evaluation (Pawson & Tilley, 1997), theory-based evaluation (Connell, 1995) and utilisation-focused evaluation (Patton, 1996). Each approach has a distinct conceptual framework. For example, realistic evaluation advocates a flexible approach. Rather than posing the question 'Does an intervention work?', realistic evaluation asks what aspects of programmes work under what conditions, when enacted by whom and at what times. The theory of change approach has been widely applied to the evaluation of early childhood interventions, including Sure Start. The hypotheses for such evaluations are based on the belief that interventions are inevitably based on theories (either explicit or implicit) about what is likely to work, but that these theories may be modified as evidence and experience unfolds.

An example of this flexibility is given by Love *et al*'s (2002) account of the evaluation of Early Head Start (p73). The conceptual framework of the programmes was based on the premise that a key indicator of success was enhanced parent–child relationships. Researchers found that changes in welfare policy increased the need for parents to use out-of-home childcare. This was at odds with the original framework of the evaluation of promoting good parent/child interactions at home. The new imperative for parents to use childcare meant that researchers modified their evaluation strategies to include visiting informal childcare providers and daycare settings. They investigated parents' views on appropriate childcare and the quality of experiences of the children in different care settings.

Even where an outcome is measurable in an evaluation it has been difficult to associate it with the processes or services designed to achieve it. Data collected about interventions, services or treatments are often drawn from qualitative sources using techniques like observation, interviews and descriptive data from case studies. The contexts or demographic characteristics in which treatments are delivered affect the programmes' efficacy in reaching populations and sustaining service use. Moreover, it is likely that there will be complex variables in the implementation of any large-scale social intervention programme that make qualitative material dense and difficult to analyse. In evaluating the success or failure of such programmes, these factors may prevent researchers establishing any secure link between the processes of the proficiency of implementation and the effectiveness of the intervention measured in terms of child and parent outcomes. This was a methodological challenge that the research team of the National Evaluation of Sure Start local programmes had to address. This article focuses on how this challenge was met.

Sure Start local programmes

Sure Start local programmes (SSLPs) were established in 1999 as the New Labour flagship anti-poverty initiative with the aim of reducing child poverty and social exclusion in many of the most disadvantaged areas of the UK (Glass, 1999). The intervention was designed to be a comprehensive, non-stigmatised, community-based programme based on identifying the needs and preferences of local communities for services. SSLPs aimed to improve children's social and emotional development, health and learning between the ages of birth and four years, to improve parental health and employability, and to strengthen families and communities.

The plan for SSLPs emerged from the Treasury[1] as the result of a cross-departmental review of services for under-fives and the evidence for their impact. It was part of the Labour government's war on child poverty, which is high in the UK, with more than 21% of children in this age group living in households with less than 60% of the national average income (ODPM, 2004). High levels of funding (around £1 million revenue per

annum plus a capital sum of at least £750,000 for building development) were awarded to deprived neighbourhoods with 800–1000 young children. The programme neighbourhoods (more than 500) did not necessarily fit within existing administrative boundaries: the idea was that they should be communities that made sense to the people who lived in them.

The community is significant to programme delivery. SSLPs use an implicit ecological model of child development, moving outward from the child through the family and community to services delivered in a joined-up and responsive way (Bronfenbrenner, 1979). The hypothesis is that the improvement of services will lead to enhanced functioning in children, parents, families and ultimately whole communities. Services are not targeted at particular families but available to all who live in the neighbourhood; thus, Sure Start had the intention to treat the whole population of children aged under four years in a boundaried area. This important principle underpinned the design of the methodology used by the National Evaluation of Sure Start.

National Evaluation of Sure Start

The National Evaluation of Sure Start (NESS) was commissioned by the UK Government in 2000 soon after the first Sure Start local programmes were set up. In setting up the tender for the evaluation the Government ruled out a randomised controlled trial (RCT), the favoured tool of a positivist approach. The brief for evaluators stipulated that the evaluation should include a longitudinal study of children from Sure Start areas (in NESS using quantitative data from standardised instruments), a cost-benefit analysis and an implementation study (in NESS using surveys, observations and quantitative data specifying change over time and case studies). (For a discussion of the consequences of the lack of RCT evidence in NESS see Rutter, 2006.)

The evaluation was one of the largest social research studies conducted in the UK, with a total budget of just under £20 million. The team included experienced researchers from the disciplines of developmental psychology, education, sociology, social work, geography, economics and medicine, as well as statisticians. It was managed by the Institute for the Study of Children, Families and Social Issues based at Birkbeck, the University of London. The scale and duration of funding offered the NESS multidisciplinary research team scope to explore innovative methodologies combining the analysis of qualitative and quantitative data.

NESS had a modular design, reflecting the SSLP holistic approach to enhancing the life trajectories of children, families and communities in SSLP areas. The five modules of the NESS research design were as follows:

(1) **The Impact Module** included cross-sectional and longitudinal studies. During 2003 the cross-sectional study aimed to recruit 12,000 nine-month-olds and 3,000 three-year-olds and their families from 150 SSLP areas. For comparative purposes 1,250 families of nine-month-olds and 1,250 families of three-year-olds were recruited in 50 areas where SSLPs were to be introduced. Families were randomly sampled within areas through child benefit[2] records. Extensive data to assess child development and family functioning were collected from the families by trained researchers during home visits lasting around 90 minutes. For nine-month-olds data collection was subcontracted to the Office for National Statistics. For the visits to three-year-olds, which included administering standardised tests to the children, a team of researchers was trained by the NESS in-house team. The measures used for the Impact Module are set out in Appendix 1 (NESS, 2005a).

(2) **The Implementation Module** aimed to provide an overview of the processes of programme implementation. Three national surveys for rounds 1 and 2 SSLPs and two applications for rounds 3 and 4 were administered annually between 2001 and 2004. Sixteen SSLPs from rounds 1 and 2 were studied in-depth as case studies. There were also a series of themed studies of cross-cutting issues such as the employability of parents, maternity services, buildings and the quality of early learning, play and childcare (NESS, 2005b).

(3) **The Cost-Effectiveness Module** analysed evidence of annual expenditure on staffing and services, the scale of resources used to deliver services in SSLPs and variability per child aged 0–4 years and patterns of spend on different services in order to explore the cost-effectiveness of the SSLPs (Final report pending).

(4) **The Local Context Analysis Module** described the SSLP areas and documented changes in records of crime, employment, health, education and regeneration at community levels over six successive years from 2000 (Barnes *et al*, 2007).

(5) **Support for local programmes:** all SSLPs were required to conduct or commission evaluations of aspects of their own work. This module

provided support to SSLPs for designing studies and using evaluation evidence. Although the costs of NESS were large, the total, cumulative amount spent on local evaluations over the years was on a similar scale.

Cumulatively, the five modules investigated the impact of SSLPs on a range of child and parent outcomes, the implementation of the SSLP intervention model, changes in the area characteristics of these neighbourhood programmes, the cost-effectiveness of the programmes and insights gained from local evaluations. It was the integrated nature of the overall research design, in particular the combination of qualitative and quantitative data from the implementation and impact modules, that made it possible to develop the methodology described in this article.

The Programme Variability Project

By 2005 it had become clear that results from the cross-sectional study within the Impact Module were disappointing but not entirely unexpected (NESS, 2005a). On average NESS could find little evidence of differences between the SSLP child, parent and family outcomes in the SSLP areas and those in the comparison communities. However, there were marked differences associated with particular programmes, with some SSLPs having distinctly better outcomes than others. For example, there was some evidence that programmes led by local authorities were associated with poorer outcomes, especially when compared to programmes led by health agencies. Possible explanations were that health authorities already had infrastructures in place for work with families with very young children through statutory midwifery and health visiting services, and experience of home-visiting in areas defined as disadvantaged. Also, where health authorities were centrally involved in SSLPs it was apparent that staff were more likely to have access to their databases of births and young children in the area (see NESS 2005a for more detailed discussion of these findings). Such variations encouraged the team to turn its attention to the hypothesis that differences in the way programmes were implemented could account for their differential effectiveness. The result was the Programme Variability Project, which bridged the NESS Impact and Implementation Modules and addressed the core question 'Why are some SSLPs more effective in achieving positive outcomes than others?'

The challenge faced has bedevilled other research teams exploring relationships between programme processes and a range of outcome measures. Although SSLPs had been set common goals, the means by which they set about achieving them were hugely variable. The variability between SSLPs was compounded by the central government requirement that communities should be centrally involved in decisions about setting up or re-shaping service provision in their localities. SSLPs were run by partnerships consisting of a combination of local statutory agencies, voluntary organisations and local people. Central government funding was administered by boards representing these partnerships, and as a result decision-making was autonomous and resulted in distinct, local approaches that were highly variable.

The one constant was the guidance issued by the Department for Education and Employment that programmes were expected to follow (DfEE, 1999–2002). This was based on a synthesis of research in the field of early childhood and family interventions conducted for the Cross-Departmental Spending Review, which had started the whole Sure Start initiative (Glass, 1999). The guidance was therefore the conceptual/theoretical framework we used as the basis for creating an instrument for measuring variations in SSLP proficiency. A concise and conceptually-based set of dimensions of programme proficiency (and potential effectiveness) using quantitative ratings was devised. The resulting instrument was applied to large amounts of qualitative data on 150 SSLPs, which were systematically collated, analysed and synthesised in order to ensure methodological rigour.

Pilot study

A pilot exercise was conducted to test the viability of the approach. Researchers from the Impact Module selected 26 SSLPs from the 150 that they were studying. Thirteen were scoring high and 13 low on two 9-month-old parenting impact outcomes (maternal acceptance and household chaos), three 3-year-old child development outcomes (verbal ability, non-verbal ability and social competence) and three 3-year-old parenting outcomes (maternal acceptance, negative parenting and home learning environment). (For an explanation of why these outcome variables were selected by the impact team see NESS 2005c, p16.)

The impact outcomes were kept from the programme variability research team so that our rating of the *likely* effectiveness of SSLPs was blind to

actual programme effectiveness. Drawing on as much data as we could gather – from government sources as well as the considerable amounts held by NESS – we drew up a matrix of some key dimensions of proficiency in programme implementation. We trawled the data for evidence of what SSLPs were doing under each dimension, and used this to make predictions about which were likely to be achieving positive outcomes and which were not. After rating the 13 most and least obviously proficient programmes on the basis of the evidence we had, it was found that 12 were linked accurately to higher and lower impact outcomes. In other words, programme proficiency seemed to be associated with programme effectiveness. The programme we had not predicted accurately was in London. In general we found London SSLPs more difficult to rate, probably because their populations were so mixed in terms of socio-economic factors. It was also more taxing to predict the effectiveness of programmes that we rated as average. However, the predictions from the pilot study were good enough, at 12 out of 13 possible correct predictions, to encourage us to proceed from the pilot phase to refine our methodology into a more systematic approach and design an instrument to measure variations in programme proficiency.

Methodology

The aim of the study was to address the key question 'Why are some SSLPs more effective in achieving outcomes than others?' The programme variability study had four phases.

(1) Producing and collating common sets of data for each of 150 SSLPs on proficiency of implementation.
(2) Rating the 150 SSLPs on dimensions of implementation proficiency.
(3) Exploring relationships between numbers of services in health, early learning/play and childcare and family support offered by SSLPs and numbers of staffing and impact.
(4) Determining the relationship between programme proficiency and their likely effectiveness as measured by child and parent outcomes.

In this article, where the focus is on the design of the Programme Variability Ratings Scale (PVRS), we do not discuss the third phase, but relationships between services and outcomes are reported elsewhere (NESS, 2005a) and work on this aspect of the evaluation is ongoing and will be reported later.

Phase (1) Producing standard data for rating the proficiency of SSLPs

The programme variability team identified 18 dimensions of implementation proficiency, based on research evidence of what works and reinforced by Sure Start programme guidance, which were likely to predict effectiveness and about which evidence could be extracted reliably from NESS and Sure Start Unit data sources. A template was designed for researchers to collate the evidence to common specifications to be used for rating SSLPs on the 18 dimensions (NESS, 2005c).

Researchers were trained in assembling a common data set for each of the 150 SSLPs included in the Impact Module of NESS. This data was extracted from existing datasets and documentation produced by SSLPs. Thus, comparable datasets for each of the 150 SSLPs in the Impact study were collated. Sources of data for synthesis within the common framework included:

- SSLP delivery plans (drawn up by local partnerships in accordance with DfES Guidance and the basis on which funding was granted to them)
- completed questionnaires from the National Survey administered by the NESS Implementation Module
- case study data where available from case studies and themed studies conducted by the NESS Implementation Module
- publications and publicity materials produced by SSLPs obtained by the NESS Implementation Module
- organisational diagram of SSLP obtained by the NESS Implementation Module
- identification of programme types based on analysis of community-level indicators carried out by the NESS Local Context Analysis Module
- local evaluation reports and materials collected by the NESS Support for Local Programme Evaluations Module
- data on SSLP progress with local evaluations from the NESS Support Module
- data on SSLP expenditure on evaluations obtained by the NESS Cost-Effectiveness Module
- quarterly returns submitted by SSLPs to the Sure Start Unit containing quantitative monitoring data on the numbers of families using each local programme.

Where no national surveys had been completed (by 19 of the 150 SSLPs), and where SSLPs had submitted

a national survey in 2002 only (six cases), sections of the national survey were re-applied using telephone interviews. More specifically, an abbreviated version of the national survey questionnaire was developed to cover the data areas considered central to the programme variability research. The questionnaire was sent to the SSLPs in preparation for the telephone interviews (NESS, 2005c).

Telephone surveys probing the views of key stakeholders in each of the 150 SSLP areas of the SSLPs' proficiency and likely effectiveness were conducted to a standard format. The views of the key informants as expert 'outsiders' to the SSLPs added further insights into the proficiency and likely effectiveness of programmes (NESS, 2005c). The expert informants included:

- programme development officers (members of regional teams, employed by Sure Start, with responsibility for overseeing and supporting SSLPs)
- Chairs and members of SSLP management boards (in post at the time of the study)
- Early Years officers (employed by the local authority and therefore likely to be knowledgeable about statutory sector involvement in the SSLPs)
- regional support staff (NESS) – (employed to offer support for local evaluations and with a knowledge of the SSLPs derived from regular visits).

Phase (2) Refining scales for rating the proficiency of SSLP implementation

The 18 dimensions of implementation proficiency were developed to be measured by a 7-point rating scale, similar to those used for established and robust measures of the quality of environment for the education and care of young children with which we were familiar (eg. the Early Childhood Environment Rating Scale (ECERS): Harms *et al*, 1998). Where a programme was rated more highly then it is regarded as being more proficient in that dimension. Each dimension was illustrated by a statement of proficiency (***Box 1***). Raters were instructed to: 'Rate each item after following the guidance notes carefully' on a scale of 1–7 as follows: 1 (Inadequate); 2, 3 (Minimal); 4 (Satisfactory); 5 (Good); 6, 7 (Excellent). As indicated already, it was an important principle for the research design that raters were operating 'blind' to the effectiveness measures, that is the Impact Module child and parent outcomes for the SSLPs.

Box 1 Dimensions and statements of proficiency for the PVRS

1.	**Vision** SSLP has a well-articulated vision that is relevant to the community.
2.	**Partnership composition** SSLP partnership board includes a balanced representation of local organisations, local education authority, social services, local NHS, voluntary and community organisations and local parents.
3.	**Partnership functioning** The partnership is functional to a high degree.
4.	**Empowerment** SSLP has the intention of creating the environment to empower users and service providers.
5.	**Communications** Communication systems reflect and respect the characteristics and languages of the host communities.
6.	**Leadership** SSLP has effective leadership/management.
7.	**Multi-agency working** Multi-agency teamwork is well established in the SSLP.
8.	**Service access** There are clear pathways for users to follow in accessing specialist services.
9.	**Staff turnover** Staff turnover is low.
10.	**Evaluation use** SSLP takes account of and acts upon evaluation findings.
11.	**Identifying users** SSLP has strategies for identifying users.
12.	**Reach** SSLP is showing a realistic and improving reach of children in the area.
13.	**Reach improvement** SSLP has strategies to improve and sustain use of services over time.
14.	**Service quantity** Service delivery reflects the guidance requirements for the provision of core services in support, health, play, early learning and childcare.
15.	**Service delivery** SSLP service delivery reflects a balance between a focus on children, family and the community.
16.	**Service innovation** SSLP shows innovative features in service delivery.
17.	**Service flexibility** Services accommodate the needs/preferences of a wide range of users.
18.	**Ethos** Overall the SSLP has a welcoming and inclusive ethos.

In ***Table 1*** below, the rating system and guidance notes for two of the 18 dimensions – 2 (Partnership composition) and 4 (Empowerment) – are given as illustrations of the processes of rating. (A complete version of this instrument and guidance notes for applying it can be found as an appendix in NESS, 2005c.) Each higher level of rating on the 7-point scale indicates an advance in both proficiency and sophistication of implementation; therefore the scales are cumulative.

Table 1 Extracts from guidance notes for applying the PVRS

2. SSLP partnership board includes a balanced representation of local organisations, local education authority, social services, local NHS, voluntary and community organisations, and local parents.

Guidance

This question relates to the relative distribution of representatives on the partnership board, and their level of seniority within their organisations (if this information is available). It also looks at efforts made to reflect the make-up of the local community within the board, as well as whether arrangements are made to make parent involvement possible (eg. training, crèche). Please make sure to write down the number of representatives from each area (health, education, etc).

The seven point scale

1) No evidence of balance in board membership/references to board.

2) Board has two or more significant gaps.

3) Board with one significant gap (eg. only one voluntary agency, no parents, no health).

4) Board includes balanced representation.

5) As 4) plus training for parents to participate/contribute to board decision-making.

6) As 5) plus evidence of an effort made to reflect the make-up of the local community (ethnicity, gender, etc) within the board.

7) As 6) plus explicit statements of value of board (eg. 'great board'!) and evidence of senior representation from agencies.

4. SSLP has an intention to empower users and service providers.

Guidance

This question focuses on efforts made by the SSLP to involve users in the running of the SSLP, and provide opportunities for development to service providers. Things that may be noteworthy are the balance between volunteers and paid staff; are parents involved in decision-making; are there exit strategies for users, services run by users, away days, staff development opportunities (including community development training, evidence of mutual respect, etc). Note that you would find evidence of community development training in the National Survey, section 3.5, under 'other'.

The seven point scale

1) No sense that users are involved at all in service planning or delivery; over-professionalisation of staffing (eg. over-dominance of highly-qualified professionals such as clinical psychologists, speech and language therapists).

2) Token mention of parents but services dominated by professionals.

3) Parents involved in some voluntary work; users on board.

4) Shows evidence of moving towards blurring the distinction between staff and users and working towards balance of voluntary and paid staff; community volunteers provide support for families; training also offered to volunteers.

5) Has a balance of voluntary and paid staff; clearly defined exit strategies for users; built-in features to develop local people's involvement; services include self-help groups or other services run by users.

6) Has whole programme away days; staff development; SSLP includes services for additional community groups (eg. grandparents, prisoners, teenagers); there is community development training for staff.

7) Shows evidence that staff are part of a learning community (eg. there are opportunities for change in staff roles and responsibilities, access to professional development); evidence of mutual respect for contributions of all parties.

Table 2 shows *the evidence* on empowerment for a programme that was rated 7 (excellent) on dimension 4 'Empowerment'. The sources indicated in the column on the left reflect those itemised in detail earlier in the article. For example, the 'Sure Start website' refers to each SSLP's public statements of intent on their websites, 'Evaluation update' refers to their local evaluation, and 'NS' refers to one of the four Implementation Module national surveys of all SSLPs, in this case its second application.

Table 2 Evidence that SSLP has an intention to empower users and service providers

Source	Evidence
Sure Start website	'Families with young children are actively engaged in planning and developing.'
Sure Start website	'Within a range of culturally sensitive services parents are given opportunities and encouragement to further develop their skills and confidence.'
Sure Start website	'...a range of workshops and training programmes for parents and early years staff to raise awareness of language development and communication.'
Delivery plan (p15)	'Parent representation at all levels will be a key feature of Sure Start... Participating parents will receive appropriate training and support and the process will be empowering and inclusive.'
Delivery plan (p17)	'It is particularly important that we tap the energies, imaginations and talents of the most excluded groups, of which there are many in....'
Delivery plan (p17)	'We must commit resources to training and developing the skills and capacities of local parents so that they can have a meaningful role in directing local services for families with young children.'
Delivery plan (p19)	Will train staff in all local services in community participation.
Delivery plan (p8)	Provide induction and equal opportunity training for all new partners.
Evaluation update (NESS)	Has formal and informal training sessions for parents only and sharing with staff and volunteers.
Evaluation update (NESS)	Group of ten parents trained to do evaluation and 10-year follow-up of planning and development exercise.
NS2 (p3)	Four FTE (full-time equivalent) volunteers doing outreach and home visiting (do general programme contact, two support families, one play, learning and childcare).
NS2 (p7)	Equipment loan scheme run by parent.
NS2 (p32)	Have parent forum.
NS2 (p33)	Has childcare, confidence building activities and training opportunities to allow parents to attend partnership meetings.
NS2 (p42)	Has training for staff and volunteers, most done separately except induction, health and safety and play and learning techniques.
NS2 (p43)	Parents involved in all aspects of staff recruitment.

Table 3 Evidence that SSLP has an intention to empower users and service providers

Source	Evidence
Publicity	Encourages dads and granddads with fathers' baby massage group and special page in newsletter.
NS2	Outreach delivered by Sure Start staff and Home-Start volunteers.
NS2	Provides career, education and training advice for parents.
NS2	Childcare provision, confidence building training and pre-meeting debrief for parents on board. Parent involvement worker recruits members to parent forum, which elects reps to board.
NS2	Training for staff not available for parents.
Newsletter	Have a volunteer day.

Another programme was rated lower at 3 (minimal) on the rating scale for the same dimension (4 Empowerment), based on the evidence given in ***Table 3***. Sources referenced in the left hand column of this table refer to leaflets for publicity, the second application of the Implementation Module national survey and a newsletter for parents.

As a second illustration, dimension 11 addressed the identification of users. Here the statement was 'SSLP has strategies for identifying users'. A 'good' SSLP (rating 5) would be one that 'identifies all potential and new users and has systems in place to identify special needs users'. Lower rated programmes would have no strategies at all, or ad hoc systems only. Higher rated programmes would have a centralised database and systematised record-keeping, routine exchanges of information between professionals about new and potential users, and regular systematic contact between SSLP staff and all families in order to identify new users as well as user needs. ***Table 4*** shows entries on the template of evidence of proficiency for an SSLP that rated highly (6) on dimension 11. The sources listed in the left hand column relate to the third application of the national survey of the Implementation Module and the interview with the local Early Years officer (EYO).

In contrast, a second extract from evidence tables (***Table 5***) is for an SSLP that was rated lower at 3 (minimal) on the same dimension, namely having strategies for identifying users. Sources refer to the first application of the Implementation Module national survey and interviews with the local authority Early Years officer and regional Sure Start Unit programme development officer.

Inter-rater reliability

Application of the rating procedure was initially carried out by four members of the programme variability research team, all of whom had detailed knowledge of the range and types of SSLPs and firsthand experience of observing the processes of programme implementation. Using the evidence

Table 4 SSLP has strategies for identifying users

Source	Evidence
NS3	SSLP uses centralised database for discovering where families live, when new babies are born and when new families move into the area. Plus multi-disciplinary team adds data directly onto SSLP database.
NS3	SSLP would expect to be informed if any children with disabilities or special needs moved to the area.
NS3	SSLP would expect to be notified of a child moving into the area registered with social services or on child protection register.
NS3	Parents/carers with special needs are identified through outreach/home visiting.
NS3	Eight out of 12 group issues identified as being significant in the area have a member of the outreach team allocated specific responsibility.
EYO interview	Good strategies in place, lots of parent involvement and community action in identifying people who need the services.

Table 5 SSLP has strategies for identifying users

Source	Evidence
NS1	SSLP discovers where new families live via information from health visitors. Discovers when new babies are born via midwifery team.
NS1	Health visitors inform SSLP when new children move into the area with disabilities or SEN health visitors monitor whether children under four are receiving routine health checks.
NS1	Systems for making contact with children not attending health checks: Health visitors send re-appointment cards and visit families to make follow-up appointments.
EYO	Feels that the geography of the area (small communities) means that mainstream services are not integrated, information is not shared and this needs improvement.
PDO	System of identification and registration of users needs tightening up.

accumulated, they all scored 42 SSLPs chosen randomly from the list of 150 SSLPs in the longitudinal study. Following this initial rating exercise, a refinement of the rating guidelines and some level descriptors was undertaken, taking into account the lessons learnt. Subsequently all the programmes were rated by two of the four original raters, both experts in the field, again without knowledge of the impact outcomes for each programme. The inter-rater reliability for these two raters was computed across all 18 dimensions. Inter-rater reliability was very good, with levels of agreement within one point on the 7-point scale being from 77% to 98% with a mean of 87%. The intra-class correlation (ie. the weighted Kappa statistic[3]) ranged from 0.55 to 0.97 with a mean of 0.77. The Spearman's rho statistic[4] ranged from 0.74 to 0.99 with a mean of 0.83. The scores from these two raters were used in subsequent analyses.

Inter-correlation of the 18 ratings

The ratings for specific dimensions of implementation proficiency for an SSLP might vary widely from each other or they might be related, in that a SSLP that scores highly on one dimension also scores highly on other dimensions. The statistical method that examines such relationships is correlation[5] and the starting point in describing the data produced by the ratings is to examine how inter-related they are through establishing the correlations (ie. statistical associations) between each possible pair of ratings.

Appendix 2 presents a table showing the intercorrelation of the 18 ratings of implementation proficiency. The ratings are all positively correlated with one another and, with a single exception, all 153 correlations are statistically significant and show a pattern of modest to strong positive correlations. In other words, programmes that scored high (or low) on one dimension tended to score similarly on others. Such a pattern of correlations indicates that there may be one or a few underlying dimensions (or factors) that are responsible for this systematic association across the 18 ratings. The statistical technique that allows the investigation of this possibility is factor analysis, which illuminates whether there are a smaller number of underlying dimensions that would capture the variation across all 18 ratings made on the 150 SSLPs and, thereby, which subsets of ratings go together to form underlying dimensions. When the 18 ratings were subjected to a factor analysis, three underlying factors emerged (the results are shown in ***Appendix 3***). Closer inspection of the factor make-up (ie. how strongly ratings align with underlying factors) reveals that all but one of the ratings ('Reach') collectively define (load heavily on) the same factor. While the three factors revealed accounted collectively for 57% of the variance in the 18 ratings across the 150 SSLPs, the first factor (on which 17 of the 18 aligned) accounted for 76% of this explained variance (ie. 42.9/56.7 = 76%). In general, then, virtually all the 18 ratings appear to be tapping into a single underlying factor reflecting general programme quality.

Analysis – Phase (4) Determining the relationship between programme proficiency and likely effectiveness

The question of whether programme variability could predict variation in programme outcomes for child development or parenting was considered for each of the three types of data separately (ratings of dimensions of proficiency; numbers of services in health, early learning/play and childcare and family support; and the numbers of staff in each), then for all three together. This article focuses on results for the first of these types of data, namely ratings of dimensions of proficiency; results of the latter two analyses are reported elsewhere (NESS, 2005a).

The focus was on child development and parenting outcomes because, first, these are the primary target for improvement in SSLPs and, second, these are the outcomes on which the cross-sectional study, comparing 150 SSLPs and 50 comparison communities, was indicating that Sure Start was having a modest influence. Two nine-month parenting outcomes were selected for analysis: maternal acceptance and household chaos. Three three-year parenting outcomes were chosen: maternal acceptance; negative parenting; and household chaos. And three child development outcomes were also chosen for three-year-olds: verbal ability; non-verbal ability; and social competence (NESS, 2005a).

The first step in the analysis addressed the question 'Do Programme Variability Ratings Overall Predict SSLP Effectiveness?' The second step addressed the question 'Do Specific Ratings Predict Specific Outcomes?' Details of the discriminant analysis[6] conducted at the first step and multiple regression[7] analysis for the second step in relation to the dimensions of proficiency, and the results, are given in the full report Variations in Sure Start Local Programme Effectiveness: Early Preliminary Findings (NESS 2005a: 16–21).

Findings

When considered collectively, the 18 ratings significantly discriminated between more and less

effective SSLPs with respect to 9-month outcomes and 36-month outcomes using complete and imputed data[8]. In order to ensure that the results were robust, discriminant analyses were repeated twice after separating the 150 sample SSLPs into two randomly selected groups of 75. The results showed correlations significantly greater than could be expected from chance. The 18 ratings are of proficiency in various aspects of the Sure Start mission, therefore this discrimination can be regarded as the result of SSLPs with higher implementation proficiency having better outcomes for children and families.

Specific ratings were found to predict one of the two nine-month parenting outcomes. *Empowerment* was found to be a significant predictor of maternal acceptance, and at three-years *identification of users* was a significant predictor for children's *non-verbal ability. Empowerment* was a significant predictor of improvements in the *home learning environment*.

There was a problematic finding, which linked *service flexibility* with lower levels of *maternal acceptance*, but there were explanations for this counter-intuitive finding in the varying degrees of inter-correlation between ratings. (For more details of findings, see NESS, 2005a; 2005c.)

Discussion

Where approaches to investigating links between outcomes and implementation of the type described in this article have been used in the past, they have been applied to interventions that have a well-defined model against which programme operations can be compared. Producing measures that could be applied across programmes exhibiting a diversity of implementation approaches had no precedent in the UK. This first application of such an innovative approach to combining quantitative (impact) and qualitative (implementation) measures has 'worked'. Use of the Programme Variability Rating Scale (PVRS) has successfully predicted overall implementation proficiency and identified some links between overall implementation proficiency and programme outcomes.

This suggests that where complex programmes have many variations in their processes they can be evaluated by the identification and articulation of key elements. Sure Start local programmes presented a plum pudding from which we have isolated 18 distinct ingredients. It may be that a refinement and reduction in these would also have provided sufficient indications of programme proficiency and effectiveness.

The principle of rendering qualitative material into quantitative measures by rating – a technique sometimes seen by practitioners as insensitive, or a 'blunt instrument' – proved capable, in this example, of detecting links between complex processes and outcomes which could not be revealed by other research techniques. An instrument like the PVRS could be designed for use in evaluating children's centres[9] at both national and local levels. It may take some time to persuade practitioners that it is in their interests that aspects of practice may be scored. But the ECERS scale referred to earlier, for example, is now widely accepted and used by students and practitioners as an instrument of choice to measure quality in early childhood settings. In applying a numerical rating exercise to what practitioners were doing in SSLPs we were able to explore possible links between professional actions and activities and their impact on children and families. A scale based on these principles, but designed for a variety of service delivery settings, could therefore be a useful tool for professionals.

The next step in developing this approach to combining quantitative and qualitative methods would be to design a way of collecting and collating datasets that provided the kind of information on services needed for rating dimensions and analysing the results, rather than super-imposing a rating system on pre-existing data, as we were obliged to do in this exercise. There is plenty of room to refine and improve the method we have shared in this article. But it will be important to continue the principles of separation that were intrinsic to the rigour of this study: separation between those delivering the intervention and those researching it, between those analysing the data and those rating it, and between those measuring inputs and those measuring outputs.

SSLPs have been re-configured and re-branded as Sure Start children's centres with different infrastructures, resources and governance. They are still complex interventions with children and families. Sure Start children's centres are designed as sites for the universal provision of services for children and their families. As with SSLPs they are not susceptible to controlled experimental approaches to evaluating what works. Children's centre managers and staff are charged with adopting a holistic approach to the families and communities they serve. The evaluation methodology described in this article has the potential to be adapted to other such interventions, although future researchers will consider themselves lucky if they have the opportunities provided for us by the extensive outcome data collected by the Impact Module of NESS. They may find instead that a methodology of this type proves a useful partner for interrogating data available from such large-scale national investigations as the Millennium Cohort Study[10] (Dex & Joshi, 2004).

Summary of policy and practice implications

- SSLPs provided an innovative model of non-stigmatised, universal, area-based early intervention in the UK.
- The imperative to work in partnership with local communities in shaping local, integrated services resulted in diversity in the programmes' responsive, flexible and innovative services.
- However, the absence of a centrally imposed curriculum and prescribed modes of service delivery presented the national evaluators of the effectiveness of SSLPs with methodological challenges.
- The Programme Variability Rating Scale was designed to investigate links between the proficiency (drawing on qualitative data common to all 150 SSLPs in the Impact study) with which SSLPs operated and their effectiveness (as measured by standardised tests in the Impact study) in terms of impact on children and parents.
- The success of the methodology indicates that it is possible to quantify qualitative datasets and subject the ratings to rigorous quantitative analyses.
- The approach will be of use as a model to inform future researchers addressing the challenges of evaluating complex social policy programmes.
- A variation of the rating scale might be usefully developed to support the managers of Sure Start children's centres in the complex task of measuring their proficiency and effectiveness.

Acknowledgements

We would like to acknowledge the work of Alistair Leyland and Helena Romaniuk on the statistical analyses for the NESS programme variability study. The authors are members of the National Evaluation of Sure Start team, based at the Institute for the Study of Children, Families and Social Issues at Birkbeck, University of London. The multidisciplinary team is led by Professor Jay Belsky (Psychology) and Professor Edward Melhuish (Psychology), and also includes Professor Jacqueline Barnes (Psychology), Professor Jane Tunstill (Social Work), Professor Sir David Hall (Paediatrics and Public Health), Dr Alastair Leyland (Statistics), Dr Martin Frost (Geography) and Pamela Meadows (Economics) www.ness.bbk.ac.uk

Address for correspondence

Professor Edward Melhuish
Institute for the Study of Children, Families and Social Issues
7 Bedford Square
London
WC1B 3RA
UK

Email: e.melhuish@bbk.ac.uk

References

Bronfenbrenner U (1979) *The Ecology of Human Development: Experiments by nature and design.* Cambridge, MA: Harvard University Press.

Barnes J, Cheng H, Howden B, Frost M, Harper G, Dave S & Finn J (2006) *Changes in the Characteristics of SSLP Areas between 2000/01 and 2003/04*. Report 016. London: HMSO.

Connell JP (Ed) (1995) *New Approaches to Evaluating Community Initiatives: Concepts, Methods and Contexts.* Queensland, MD: Aspen Institute.

Dex S & Joshi H (Eds) (2004) *Millennium Cohort Study First Survey: A users' guide to initial findings.* Available from: www.esds.ac.uk/longitudinal.

DfEE (Department for Education and Employment) (1999–2002) *Sure Start, A Guide for Trailblazers/second/third/fourth wave programmes.* London: DfEE.

Glass N (1999) Sure Start: the development of an early intervention programme for young children in the United Kingdom. *Children in Society* **13** (4) 257–264.

Harms T, Clifford R & Cryer B (1998) *Early Childhood Rating Scale Revised.* New York and London: Teachers' College Press.

Love J, Kisker EE, Ross CM, Schochet PZ, Brooks-Gunn J, Paulsell D, Boller K, Constantine J, Vogel C, Fuligini AS, Brady-Smith C (2002) *Making a Difference in the Lives of Infants and Toddlers and their Families. The Impacts of Early Head Start. Volume 1: Final Technical Report.* Princeton, NJ: Mathematica Policy Research Inc. Available from: www.mathematica-mpr.com?PDFs?ehsfinalvol1.pdf.

NESS (National Evaluation of Sure Start) (2005a) *Early Impacts of Sure Start Local Programmes on Children and Families. Report 013.* London: DfES.

NESS (2005b) *Implementing Sure Start Local Programmes: An In-Depth Study, Report 007.* London: DfES.

NESS (2005c) *Variations in Sure Start Local Programmes' Effectiveness: Early Preliminary Findings Report 014.* London: HMSO.

Office of the Deputy Prime Minister (ODPM) (2004) *The English Indices of Deprivation, 2004. Summary (revised)*. London: ODPM.

Olds DL, Henderson CR, Kitzman H, Eckenrode JJ, Cole RE & Tatelbaum RC (1999) Prenatal and infancy home visitation by nurses: Recent findings. *Future of Children* **9** 44–66.

Patten MQ (1996) *Utilization-focused Evaluation.* Beverley Hills, CA: Sage.

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

Ramey CT, Campbell FA, Burchinal M, Skinner ML, Gardner DM & Ramey SL (2000) Persistent effects of early childhood education o high-risk children and their mothers. *Applied Developmental Science* **4** 2–14.

Rutter M (2006) Is Sure Start and Effective Preventive Intervention? *Child and Adolescent Mental Health* **11** (3) 135–141.

Sanders MR (2003) Triple P-Positive Parenting Programme: A population approach to promoting competent parenting. *Australian e-Journal for the Advancement of Mental Health* **2** (3). Available from: www.auseinet.com/journal/vol2iss3/sanders.pdf.

Webster-Stratton C (1993) Strategies for helping families with oppositional defiant or conduct-disordered children: The importance of home and school collaboration. *School Psychology Review* **22** 437–457.

Endnotes

[1] The Treasury is the UK government department responsible for developing and executing the public finance and economic policy.

[2] Child Benefit is a tax-free monthly payment to anyone bringing up a child or young person. It is not affected by income or savings, so most people who are bringing up a child or young person qualify for it.

[3] The Kappa coefficient is used to judge the inter-rater reliability between two or more individuals. Thus, when two people are making judgements on the same thing, for example a questionnaire, the Kappa coefficient measures the extent to which they agree with each other (ie. give the same responses).

[4] The Spearman Rho correlation measures the magnitude and direction of the association between two variables. It does not require the variables to be measured on interval scales; it can be used for variables measured at the ordinal level.

[5] Correlation is a measure of statistical association and is measured from –1 (complete negative relationship) to +1 (complete positive relationship), with o representing no relationship.

[6] Discriminant analysis is used to determine which variables discriminate between two or more naturally occurring groups (in this case more and less effective Sure Start Local Programmes).

[7] Multiple regression is a statistical technique that predicts the variation in values of one (dependent) variable (in this case SSLP effectiveness) on the basis of two or more other (independent) variables.

[8] The effectiveness scores for SSLPs were derived from two datasets, one of which included cases for which 100% of the variables were available, the other which included all eligible individuals, even if their data was incomplete (imputed data).

[9] Children's Centres are intended to be places where children under five years old and their families can receive seamless, holistic, integrated services and information, and where they can gain access to help from multidisciplinary teams of professionals. They provide a range of services, including free advice to families on health, benefits and housing, as well as support for young children with disabilities and learning difficulties and full day care and early education. The Government is committed to every community having a Children's Centre by 2010.

[10] The Millennium Cohort Study is following a sample of 18,819 children drawn from all live births in the UK over 12 months (from 1 September 2000 in England & Wales and 1 December 2000 in Scotland & Northern Ireland).

Appendix 1 Impact study outcome measures

Child cognitive ability	
Verbal ability*	Language expression and comprehension abilities (subscale of BAS)
Non-verbal ability*	Spatial and number skills (subscale of BAS)
Social and emotional	
Child social competence*	A construct of 'Pro-social' (shows concern for others, shares, liked by others) and 'independence' (works things out for self, chooses activities or self, persists with difficult tasks)
Behaviour difficulties*	A construct of: 'Conduct Problems' (antisocial or disruptive behaviour; fights/bullies, temper tantrums, argues), 'Emotion Regulation' (worried/anxious behaviour, worries, clingy, tearful, fearful),'Hyperactivity' (rest-less, distractible, impulsive, overexcited), and 'Overall Difficulties' getting along with others, concentrating, behaving).
Physical health	
Birth weight**	Child's weight at birth in grams
Ever breastfed**	Divides the mothers in 2 categories, those that attempted breastfeeding v. those that did not
Breastfed min of 6 weeks**	Two groups; mothers that breastfed less than 6 weeks vs. those breastfeeding for longer
One or more accidents	One or more accidents in past year (9mths for 9mth olds)
Admitted to Hospital	One or more hospital admissions in past year (9mths for 9mth olds)
Parenting/family functioning	
Supportive parenting	A construct of: 'Responsivity' (observations of mother praising, responding, showing affection) and 'Acceptance' (not observing scolding/derogating, spanking, physically restraining)
Negative parenting*	A construct of: 'Parent/child conflict' (parent-child struggles, child easily with parent, conflict with discipline), 'Parent/child closeness' (affectionate relationship, child seeks comfort, child shares feelings), 'Harsh Discipline' (frequency of [reported] swearing, threatening, smacking, slapping child), and 'Home Chaos' (disorganised, noisy, lacking regular routine)
Home learning environment*	Learning opportunities provided in home; child read to, taken to library, engaged in play with letters and numbers, taught songs/rhymes
Father involvement	Looks after, feeds, plays with child (as reported by mother)
Home chaos**	Disorganised, noisy, lacking regular routine
Maternal well-being	
Malaise: depression measure	Jittery, tired, depressed (bad for parenting and child development)
Self-esteem	Positive feelings about self (good for parenting and child development)
Local area measures	
Mother's area rating	A score given by the mother, for her local area
Observer's area rating	A score for the area given by the interviewer
Services	
Total support services used	Number of different types of services from which respondent received support
Total support usefulness	Usefulness of support services used (mean score of service use)

* Denotes outcomes for the 3-year-old group only
** Denotes outcomes for the 9-month-old group only

Full references to the standardised instruments used are located in NESS Report 13 (NESS 2005a) referenced in the bibliography of this article.

Appendix 2: Intercorrelations of the 18 ratings of implementation proficiency

	Partnership –composition	Partnership -functioning	Empowerment	Communication	Leadership	Multi-agency	Pathways	Staff turnover	Use Evaluation	Identify users	Reach	Reach - strategies	Services - quantity	Services - delivery	Services - innovation	Services - flexibility	Ethos
Vision	0.37**	0.49**	0.48*	0.43**	0.40**	0.42**	0.40**	0.35**	0.32**	0.39**	0.37**	0.46**	0.41**	0.42**	0.37**	0.36**	0.41**
Partnership – composition	1	0.44**	0.50**	0.26**	0.46**	0.34**	0.08	0.19*	0.19*	0.22**	0.25*	0.29**	0.27**	0.23**	0.30**	0.32**	0.32**
Partnership – functioning		1	0.49**	0.36**	0.57**	0.48**	0.23**	0.22**	0.27**	0.34**	0.26**	0.44**	0.42**	0.40**	0.40**	0.46**	0.43**
Empowerment			1	0.50**	0.49**	0.52**	0.32**	0.22**	0.32**	0.34**	0.24**	0.40**	0.44**	0.42**	0.46**	0.43**	0.53**
Communication				1	0.49**	0.45**	0.45**	0.30**	0.28**	0.29**	0.18*	0.35**	0.48**	0.51**	0.45**	0.37**	0.54**
Leadership					1	0.59**	0.35**	0.43**	0.36**	0.45**	0.19*	0.46**	0.42**	0.39**	0.45**	0.49**	0.63**
Multi-agency						1	0.38**	0.35**	0.32**	0.30**	0.17**	0.43**	0.45**	0.38**	0.36**	0.40**	0.50**
Pathways							1	0.37**	0.34**	0.44**	0.21*	0.49**	0.33**	0.38**	0.29**	0.35**	0.46**
Staff turnover								1	0.42**	0.36**	0.27**	0.39**	0.34**	0.29**	0.26**	0.27**	0.43**
Use Evaluation									1	0.45**	0.28**	0.47**	0.39**	0.36**	0.40**	0.29**	0.46**
Identify users										1	0.33**	0.57**	0.41**	0.36**	0.46**	0.35**	0.45**
Reach											1	0.42**	0.25**	0.24**	0.29**	0.31**	0.22**
Reach – strategies												1	0.46**	0.52**	0.44**	0.44**	0.56**
Services – quantity													1	0.77**	0.47**	0.42**	0.49**
Services – delivery														1	0.51**	0.43**	0.55**
Services – innovation															1	0.55**	0.54**
Services – flexibility																1	0.54**
Ethos																	1

*p<0.05 **p<0.01

Appendix 3 Factor analysis of the 18 ratings of SSLP implementation proficiency

FACTOR STRUCTURE OF 18 RATINGS (Oblique Rotation)

	Component		
	1	2	3
Vision	**.667**	-.047	.208
Partnership composition	**.502**	-.536	.376
Partnership functioning	**.660**	-.381	.167
Empowerment	**.696**	-.363	-.006
Communication	**.664**	-.094	-.390
Leadership	**.743**	-.231	-.009
Multi-Agency	**.677**	-.226	-.128
Pathways	**.583**	.386	-.197
Staff turnover	**.538**	.351	.095
Use evaluation	**.585**	.369	.111
Identify users	**.636**	.351	.201
Reach	.438	.223	**.615**
Reach strategies	**.733**	.279	.174
Services quantity	**.711**	.051	-.278
Services delivery	**.708**	.087	-.343
Services innovation	**.688**	-.005	-.052
Services flexibility	**.667**	-.112	.019
Ethos	**.786**	.022	-.185
Eigenvalue	7.72	1.36	1.12
% variance explained	42.9	7.6	6.2

The factor loadings show that 17 of the 18 ratings load most heavily, and above 0.5, on the first component (factor). 'Reach' is the exception in that it loads most heavily on component 3, while still retaining a moderately high loading (0.438) on component 1

