Department for Education

Evaluation of MTM's Signs of Safety Pilots

Evaluation report appendices

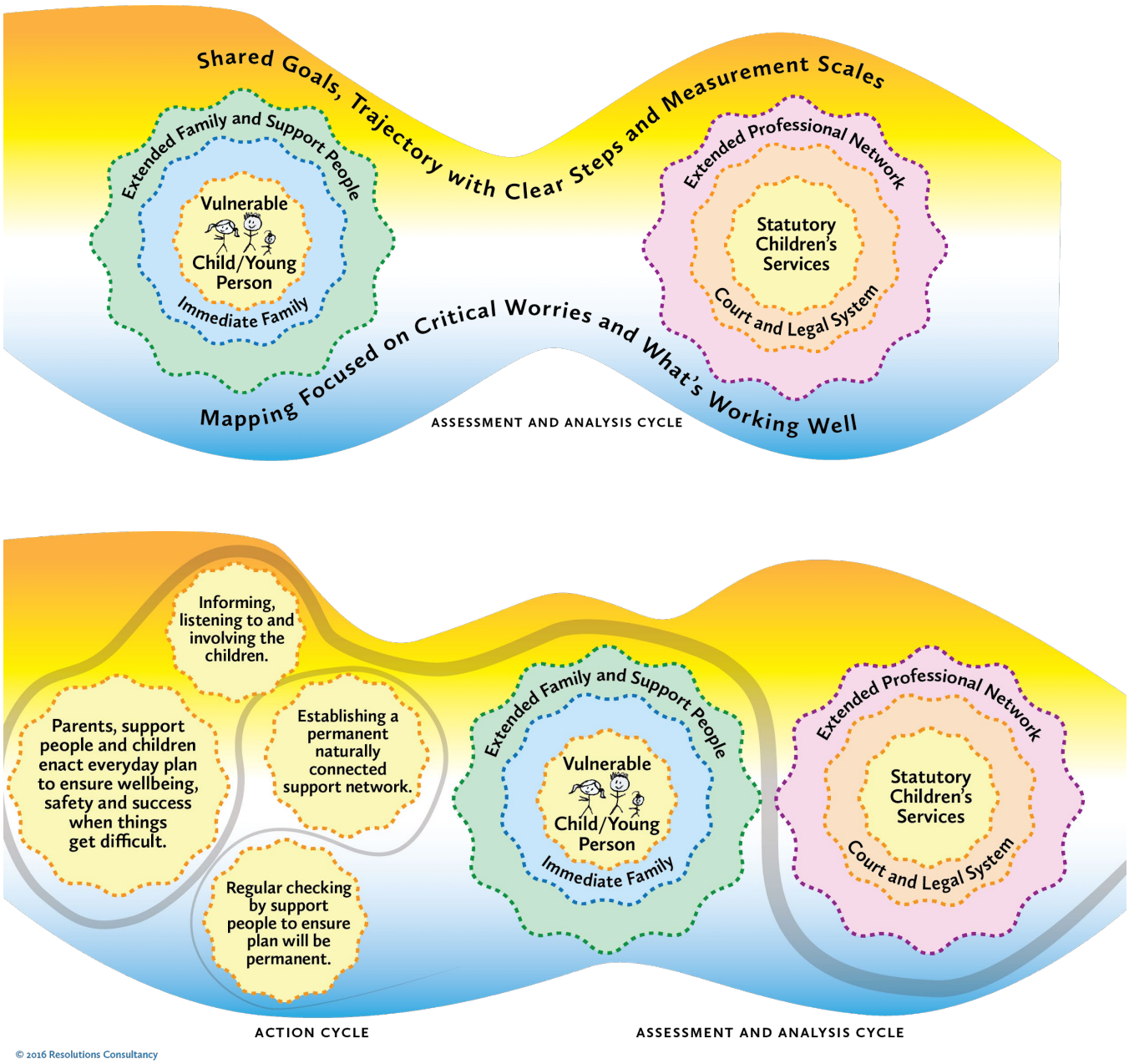
October 2020

Mary Baginsky, Ben Hickman, Jess Harris, Jill Manthorpe, Michael Sanders, Aoife O'Higgins, Eva Schoenwald and Vicky Clayton

Contents

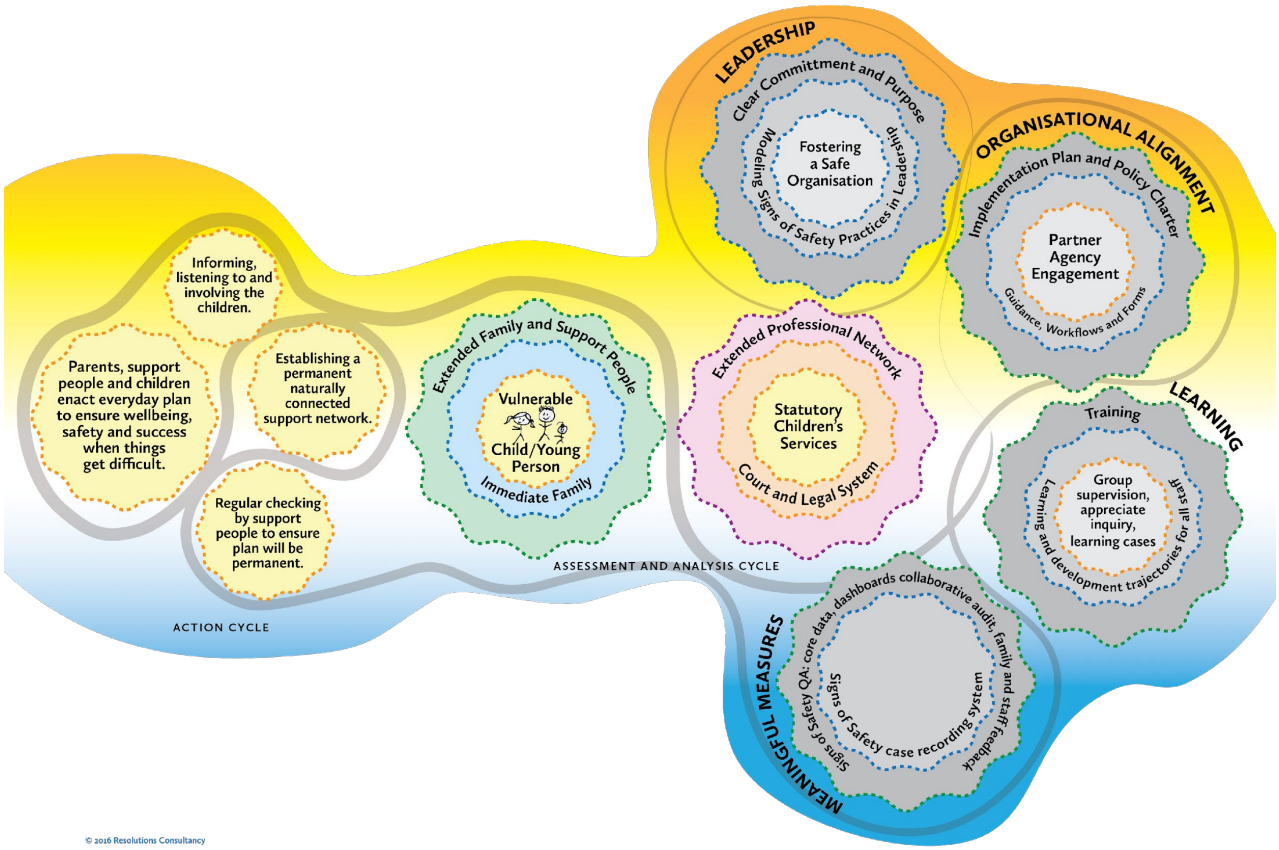
Appendix 1: Project theory of change

Figure A1.1: Signs of Safety Practice Theory of Change



© 2016 Resolutions Consultancy

Figure A1.2: Signs of Safety Organisational Theory of Change



Appendix 2: Evaluation's logic model

Table A2.1: Evaluation's logic model

Context	Elements	Mechanisms	Interventions	Outcomes
10 local authorities in England 9 had been involved in EIP 1and 1 joined in EIP2 1 that had been involved in EIP 1 subsequently left the project Ofsted ratings	Assessment and planning with the family Building a network of family and social supports Using the assessment and planning framework to focus on analysis and building the central place of the family in demonstrating safety Moving through interlinked cycles of assessment and analysis together with action to achieve the goals of the case, as assessed by the family, its network and the professionals Commitment from leadership group Support arrangements for practice Training Supervision Group supervision	Safety plans: detail actions by the family and the network members in the face of identified dangers and identified triggers (red flags) for those dangers include contingency actions should the planned actions be impeded that are written out in detail for adults that are set out in a Words and Pictures for the children (and the adults). Case work that should occur in conjunction with a safety plan: all families being encouraged, supported and expected to bring a network into the case work mapping* with the family and network including use of the analysis categories with danger statements and safety goals agreed with the family Three Houses work with the child(ren) and this being shared with the parents Words and Pictures explanation for the child(ren) being prepared with the family and this being shared by the family with the children and the network. *Mapping involves: What is worrying (past hurt, danger statement and complicating factors) What is working well (existing strengths and existing safety) What needs to happen (safety goals, friends and family safety network, managing safety plan, trajectory, bottom line and scaling)	Basic training Advanced training Coaching for practice leaders Supporting the transition of basic training in-house Developing specific areas of practice and staff groups Organisational consultancy Leadership development Organisational alignment of processes and systems, and meaningful measures (aligning quality assurance and information management)	Families and children feel more empowered, are better able to understand children's services' concerns and requirements and so are better able to address the concerns for more effective outcomes and reduced re-referrals Practitioners report greater clarity, job satisfaction and commitment leading to improved staff retention and reduced absenteeism The number of children removed from families reduces as the number of families being supported intensively increases, including greater confidence to close cases

Appendix 3: Self-profiling instrument

Self-profiling instrument: Pilot _____ (Please add authority's name)

Please indicate (with a tick or cross) where you think your authority is in relation to each of these items:

a) at the start of the Signs of Safety project
b) at the end of the Signs of Safety project.

The information you return will be held in strictest confidence and will be used as one of the many factors contributing to the evaluation of the programme. Not all of the items will be **exclusively** linked to Signs of Safety and it would be helpful if you would place an asterisk next to items that are **directly linked** with other initiatives in addition to Signs of Safety.

Please rate each on a scale of 1–10 where 1 = Not at all and 10 = Fully

Components of Signs of Safety
Organisational culture
Progress along the path towards building constructive working relationships between professionals and family members?
Creating a collaborative culture with parents around child protection practice, whilst remaining vigilant and realistic about risk.
Providing an effective early help offer to allow intervention at the right time
Spending the necessary time with adults in families
Spending the necessary time with children in families
Confidence that your service is intervening at right time
Creating a culture of practice led evidence
Creating a culture where it is permissible to admit you may have it wrong
Embedding an organisational commitment to Signs of Safety
Fostering a safe organisation - building confidence that workers will be supported through anxiety, contention and crises
Practice issues
Using plain language that can be readily understood by families
Capturing the voice of the child in safeguarding practice and direct work
Using tools to engage children and young people e.g. three houses; fairies and wizards
Using Words and Pictures explanations
Using statements focusing on specific observable behaviours
Separating fact from hearsay
Mapping cases by individual social workers
Mapping cases in teams
Managing safety plans over time linked with progression
Using safety plans across initial and review child protection conferences and in all related groups
Using safety plans developed with families
Development of appreciative inquiry work with families
Developing family networks from the outset
Placing Family Network Meetings at the centre of all processes
Allowing families to run Family Network Meetings

Learning
Basic training in Signs of Safety is provided in-house as part of compulsory introductory training
Develop in house training team to deliver SOFS basic training
Providing basic 2-day training for all social workers
Providing basic 2-day training for all other social care staff working with children
Providing advanced 5-day training for all managers
Providing advanced 5-day training for all social workers
Embedding Signs of Safety approaches and principles across all training for those working in children's social care
Processes
Aligning paperwork with Signs of Safety practice
Aligning Initial Child Protection Conferences with Signs of Safety
Aligning Review Child Protection Conferences with Signs of Safety
Achieving consistency in the quality of social work decision making and practice
Revise, negotiate and implement changes to Public Law Outline (PLO) policy, procedure and practice to fit with Signs of Safety
Aligning quality assurance and audit processes with Signs of Safety
Aligning IT systems with Signs of Safety
Structures
Achieving manageable caseloads for all practitioners
Supporting social workers with administrative tasks
Recruiting high quality staff
Retaining high quality staff
Establishing practice leadership and supervision processes to support Signs of Safety
Creating a culture of appreciative inquiry across staff interactions
Leadership and 'staying the journey'
Embedding Signs of Safety as the organising framework for all child protection practice
Planning for expected difficulties
Planning for unexpected difficulties
Having a vision which is shared and which is sustainable even if key people leave
Embedding an organisational commitment to Signs of Safety

Appendix 4: Progress on implementation of Signs of Safety

Individuals in focus groups assessed each item on a scale of 1–10 where 1 = not at all and 10 = completely

- Plans remain central to practice from assessment through to case closure
- Plans are regularly reviewed and revised
- Having sufficient time to spend with families
- Establishing naturally connected support networks with families
- Informing, involving and listening to children
- Clear commitment to Signs of Safety from management
- It feels like a safe organisation in which to work
- Family Court engaged in Signs of Safety approach
- Recording systems aligned with Signs of Safety
- Quality assurance systems aligned with Signs of Safety
- Group supervision and appreciative inquiry in place
- Partner agency involvement

Appendix 5: Contrast study

The study was located in one team in 4 authorities – 2 SofS pilots (Pilots 4 and 9) and 2 non-SofS authorities. Of the 2 non-SofS sites one had adopted a restorative practice approach across children's services and the other had developed an approach to working with families based on striving for positive change through a number of routes including motivational interviewing and multi-disciplinary teams with specialist practitioners on domestic abuse, parental mental health and alcohol/substance misuse. In Pilot 9 and in the 2 contrast sites the work took place at approximately 6-month intervals (T1, T2 and T3). Pilot 4 did not enter the contrast study until summer 2019 when it became evident that the work could not continue in Pilot 7, so the fieldwork was conducted between August 2019 and February 2020.

One team had experienced a change of manager and there was considerable staff turnover and staff shortages in 2 others, although all 4 teams were working under considerable pressure. At one point in one contrast site, a team that should have consisted of a manager and 6 social workers plus a drug worker and domestic violence worker consisted of a manager, one part-time experienced social worker, a one-year post-qualification social worker and another social worker who was in the assessed and supported year, plus the domestic violence worker.

Given the limitations of resources devoted to the evaluation that had to be spread across the 4 areas at 3 time points there was a heavy reliance on social workers' co-operation which, for the most part, was extremely good. Nonetheless, it is important to recognise that, for some social workers, however much it was stressed that individuals would not be judged, that was how it would have been perceived. In a few cases parents did not agree to participate but it was evident that in some teams the bigger challenge was the reluctance on the part of social workers to suggest participation. Understandably some cases were excluded because they were at a particularly challenging point or because a social worker said a family would never agree or that the researcher's presence would change their relationship with the family. Overall, as far as it was possible to judge, a high proportion of families who were approached agreed to take part and there was a high level of engagement in providing feedback.

All observations were conducted by the same evaluator. While this brought consistency, if there were biases there was no one else to provide challenge. While interesting data were collected it must primarily be viewed as a trial for how to conduct a subsequent study as many lessons were learnt, not least the benefits that would attach to embedding researchers in teams for a period of time while the work was concluded.

The instruments used during the contrast study were:

Organisational social context tool: The Organizational Social Context (OSC) Measure is a nationally normed (for use in 2 settings: child welfare and mental health) and psychometrically proven 105-item scale that measures the cultures and climates of child welfare and mental health organisations. It can be administered online or using paper scan forms. The OSC Measure is completed by front-line staff (rather than managers or leaders) to obtain the most representative view of an organisation's culture and climate. Reliability coefficients for OSC dimensions range from .78 to .94.

Clinical Competence-Based Behavioral Checklist (CCBC): The CCBC is a tool for assessing performance in social work practice that consists of 4 categories: (1) interviewing skills; (2) cultural empathy; (3) assessment and intervention strategies; and (4) comprehensive evaluation (see Lu et al., 2011). To increase the reliability of the measure, an 'overall score' category was added. This assesses a broader set of skills than individual categories. Regehr et al.(1999) have reported that the scores for an overall assessment category are at least as reliable as the scores for individual categories and in some cases more valid.

Families and social workers

Working Alliance Inventory (WAI)

- WAI – short version for SWs
- WAI – short version for families

See https://wai.profhorvath.com

Families

Hampshire County Council Children's Services Family Feedback

Family Chart developed at Round 1 to collect feedback on children's social care

Family Feedback Scaling Chart

Practice Elements of Signs of Safety Chart

Client Engagement in Child Protection Services (Yatchmenoff)

Social workers

Survey for attached social workers in SofS sites

Appendix 6: Views of senior leaders in the 2 pilots that exited the SofS project

Senior leaders in both pilots had joined the authorities after Round 1. The children's service department in the pilot that did not take part in Round 2 had been judged inadequate by Ofsted in April 2017 but by June 2019 was found to have improved substantially on all aspects and was then judged to be 'good'. The pilot that left the project half-way through Round 2 had been found to require improvement at the end of Round 1 but by June 2018 it received an 'inadequate' judgement. The new senior management team did not consider that SofS, in the form that was advocated during EIP, fitted with the improvement journey the authority was following.

Both inspection reports that led to the 'inadequate' judgements had highlighted the fact that children were being left at 'risk of significant harm'. While there were several contributory factors in both departments, the question of how SofS dealt with risk was at the heart of the decision to move away from it.

The discussion in both former pilots was very similar. SofS was seen to have become a substitute for basic social work skills and the '3 columns'[1] was seen as the assessment rather than a tool to aid analysis. It was being used superficially and as a result was leading to risky practice because it encouraged an overly strengths-based approach with a tendency to minimise risk. There was an over-reliance on parental self-reporting and the production of plans was based on what families said they would do rather than on an understanding of what was required, what was needed to achieve change and an assessment of whether that change was sustainable:

> There is a basic flaw with it, which is that you rely on people who may not be able to do the right thing, to do the right thing and in so doing it minimises professional judgement.
>
> The approach facilitated the 'rule of optimism' which, in the (authority's) context was at best superficial and at worst meant risk was missed.

It was said to have been used as a way of processing cases more quickly to deal with the demand and to have been imposed from above, depending on a small number of

[1] Part of Signs of Safety assessment and planning framework that identifies what children's services are worried about (past harm, future danger, complicating factors), what is working well (existing strengths and existing safety) and what needs to happen (family and child protection authority safety goals and next steps for future safety)

advocates, rather than implemented from the bottom, taking staff along and listening to their response.

Alongside these concerns was one focusing on the electronic recording systems that had evolved to reflect SofS but which, in the opinion of these informants, did not reflect the statutory framework within which work must take place. It was reported to fail to reflect the basic statutory functions in terms of information sharing, chronology, history as an indicator for the future, and arriving at clear conclusions about risk and strength that are based on balance.

Restorative practice had been introduced into both former pilots. One informant summed up the reasons for adopting this approach:

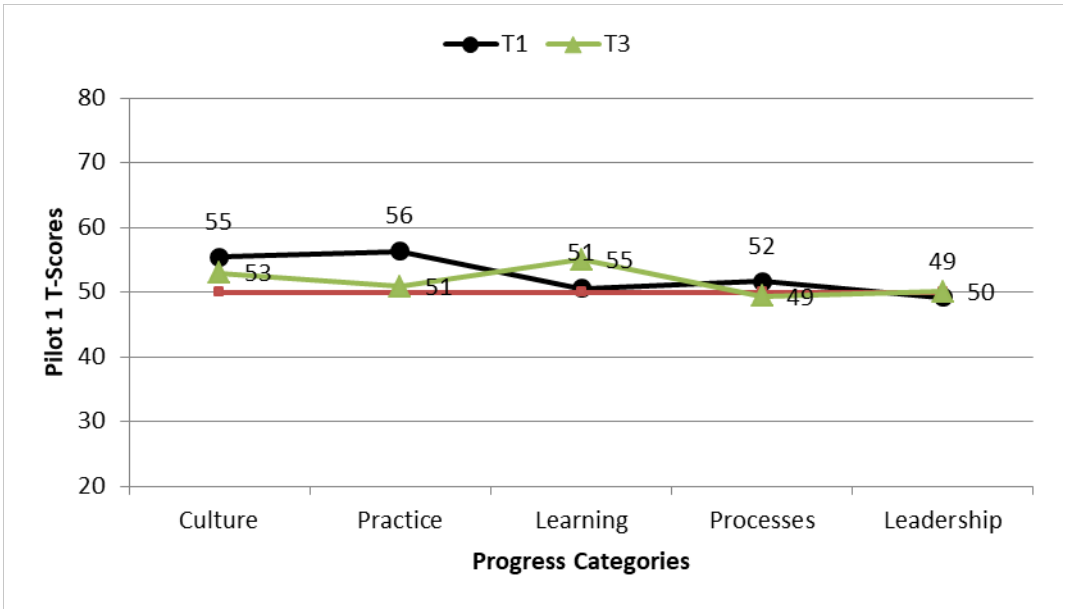
> Restorative, from my point of view, provides an overarching umbrella around some ways of being; it's more about language and approach. It helps with interventions, it helps with behaviours and culture, from my point of view. What it gives people, I think, is an ability in equal measure to support and challenge; very, very simple and what it does, I think, is provide a focus on expectation and help. So where it's really helpful from a safeguarding point of view is when you're approaching a family in need of help, as most of our families are, but where you also have to shift up a gear quite quickly, it helps staff to think about how we do this without alienating and causing conflict … And it easily over-layers onto the statutory framework.

Appendix 7: Profiles

The profiling exercise asked strategic leads in the 9 pilots to rate their progress (1–10) at the beginning (T1) and the end (T3) of Round 2 on 50 items organised into 5 categories: organisational culture; practice issues; learning; processes and leadership. All pilot scores were converted to T-scores based on the mean average response of all pilots. A score of 50 represents the mean and a difference of 10 from the mean indicated a difference of one standard deviation.

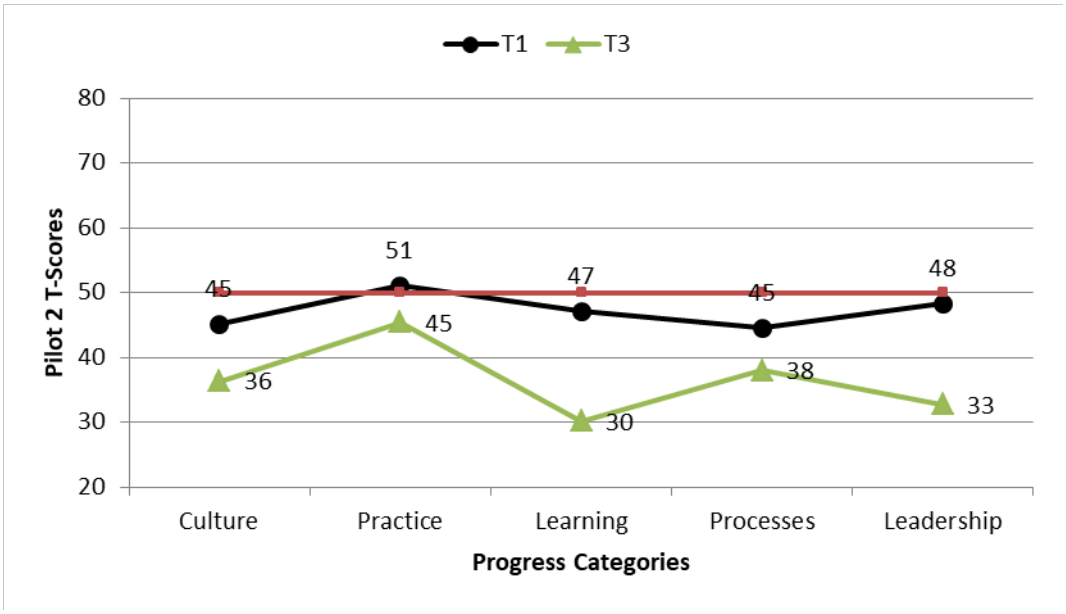
Pilot 1 had slightly higher than average scores for culture and practice at T1, with average scores for the remaining categories. At T2 its scores had decreased slightly in both culture and practice, although they remained just above average, and had improved in learning.

Figure A7.1: Pilot 1 T-Scores for profile of SofS progress at T1 and T3



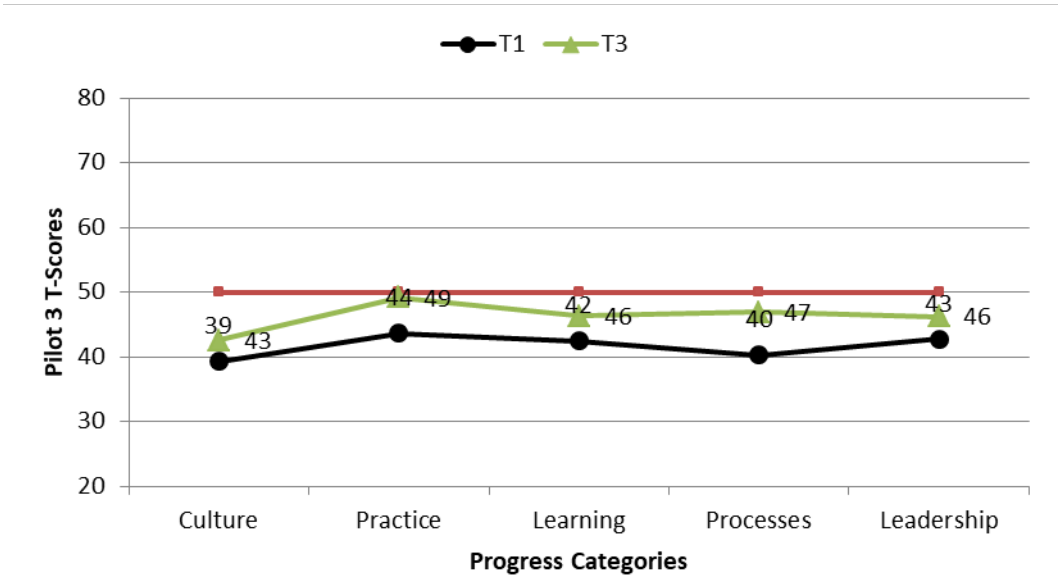
At T1, Pilot 2 was slightly below average in each of the categories other than practice (where it was just above average). By T2, its scores had decreased in every category and were one standard deviation below the average in culture, processes and leadership, and 2 standard deviations below in learning.

Figure A7.2: Pilot 2 T-Scores for profile of SofS progress at T1 and T3



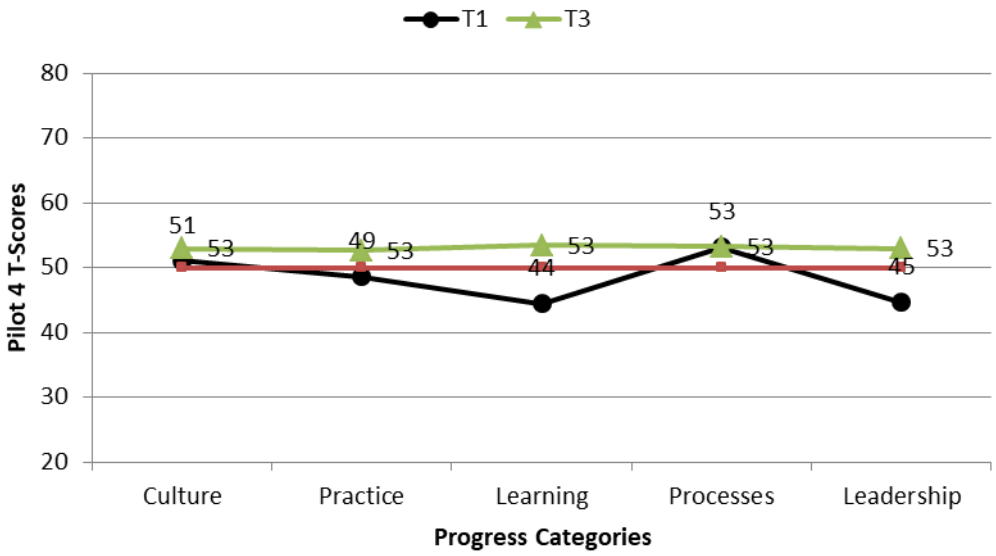
Pilot 3 was below average in all of the categories at T1 and was one standard deviation below in culture and processes. By T2 it had improved in each of the categories but remained below average in all.

Figure A7.3: Pilot 3 T-Scores for profile of SofS progress at T1 and T3



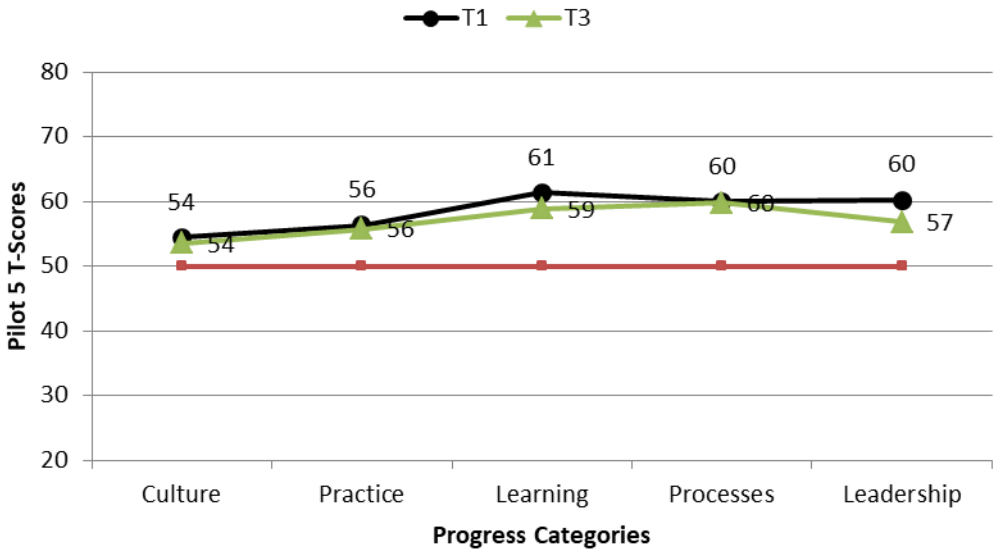
At T1, Pilot 4 was below average in learning and leadership and above average in processes. It had improved in each of the categories other than processes by T2, to become slightly above average in all categories.

Figure A7.4: Pilot 4 T-Scores for profile of SofS progress at T1 and T3



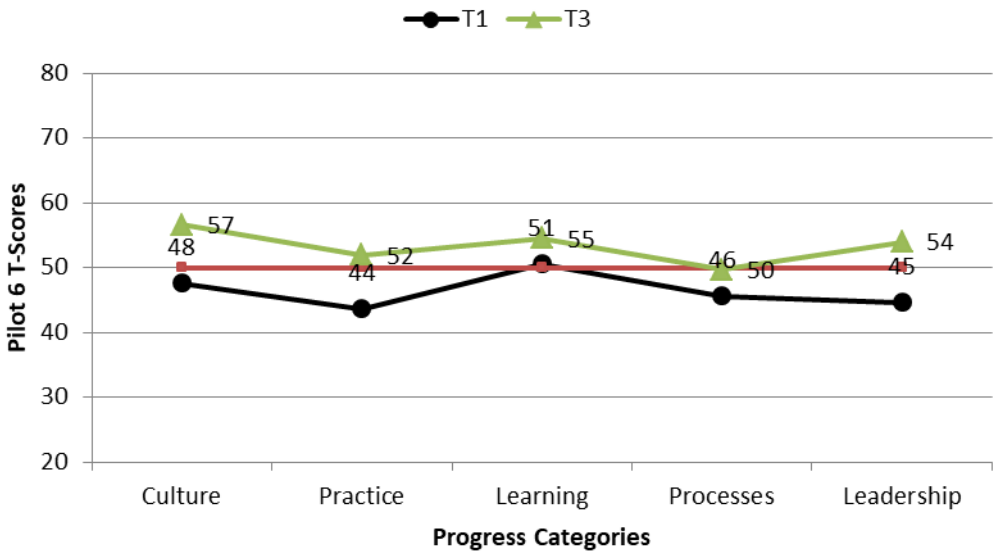
Pilot 5 was above average in all categories at T1, and was one standard deviation above in learning, processes and leadership. It remained above average in all categories at T2, although only processes remained one standard deviation above the average.

Figure A7.5: Pilot 5 T-Scores for profile of SofS progress at T1 and T3



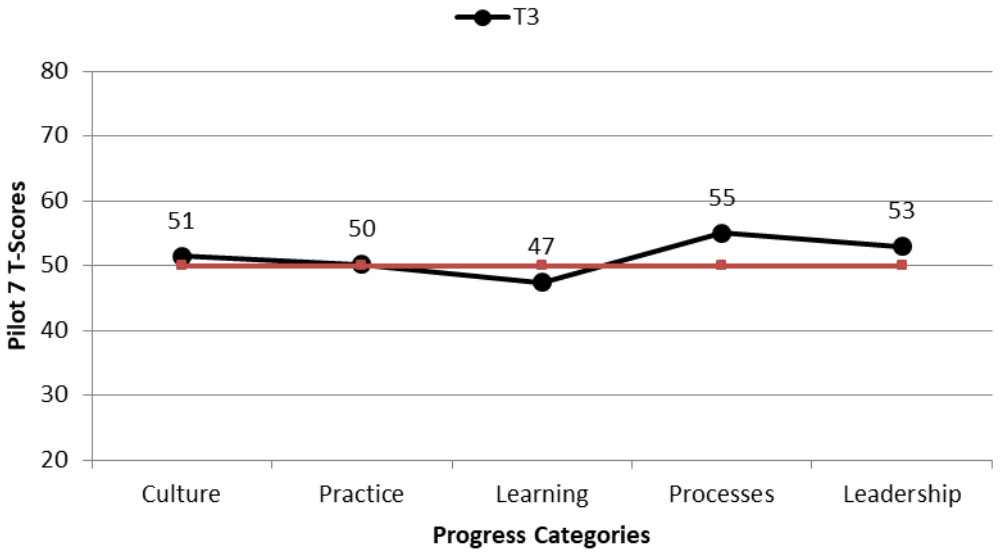
Pilot 6 was slightly below average in all categories other than learning at T1. By T2 it had improved in every category to become slightly above average in all categories other than processes, where it remained just below average.

Figure A7.6: Pilot 6 T-Scores for profile of SofS progress at T1 and T3



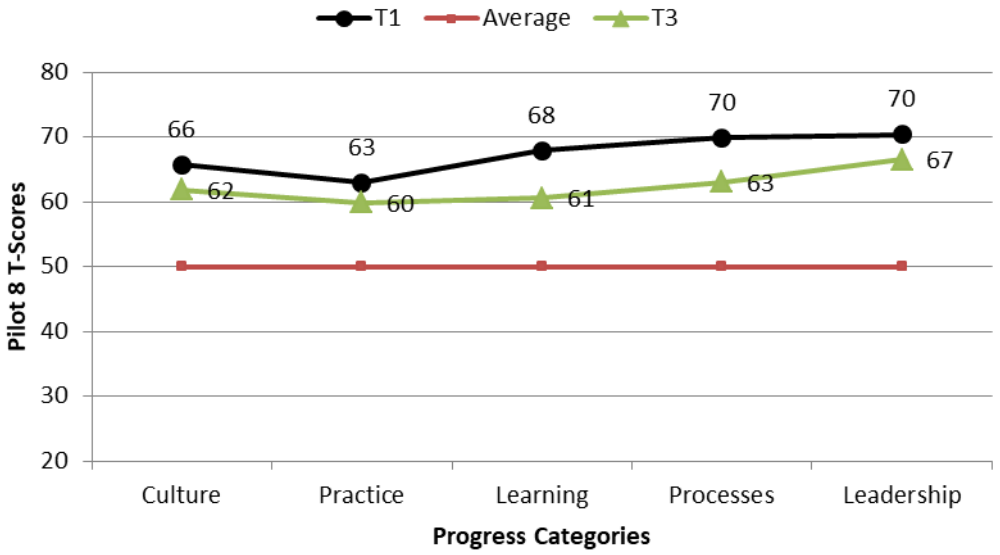
No data was available for Pilot 7 at T1. At T2 it was slightly below average in learning and slightly above average in processes and leadership.

Figure A7.7: Pilot 7 T-Scores for profile of SofS progress at T1 and T3



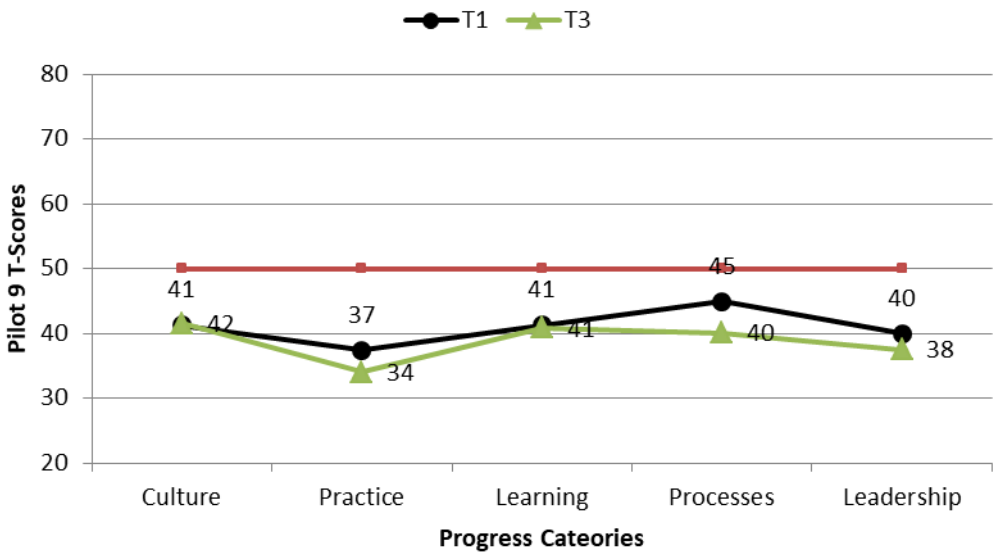
Pilot 8 was at least one standard deviation above average in all categories, and was 2 standard deviations above average in processes and leadership. It remained one standard deviation above in every category at T2, although scores had fallen in every category.

Figure A7.8: Pilot 8 T-Scores for profile of SofS progress at T1 and T3



Pilot 9 was below average in each category at T1, and was one standard deviation below in practice and leadership. By T2 it had fallen in each category other than culture (where it remained below average), and was one standard deviation below average in practice, processes and leadership.

Figure A7.9: Pilot 9 T-Scores for profile of SofS progress at T1 and T3



Appendix 8: Focus group profile data summary table

Table A8.1 Focus group profile data summary

Elements of SofS	P9	P9	P9	P7	P7	P7	P8	P8	P8	P2	P2	P2	P1	P1	P1
P = Pilot no. N = Now F = Future	T1 N	T1 F	T2 N	T1 N	T1 F	T2 N	T1 N	T1 F	T2 N	T1 N	T1 F	T2 N	T1 N	T1 F	T2 N
Plans remain central to practice from assessment through to case closure	6.3	8.5	7.0	5.7	8.2	6.6	6.0	8.0	7.5	5.4	8.1	6.4	4.5	6.6	6.3
Plans are regularly reviewed and revised	6.5	8.5	7.3	5.8	8.5	7.1	5.7	7.4	7.7	5.2	7.9	7.0	4.8	6.0	7.2
Having sufficient time to spend with families	4.7	7.8	5.7	5.2	8.3	4.9	3.9	6.5	4.6	3.2	7.4	5.4	3.5	4.7	4.5
Establishing naturally connected support network with families	5.1	7.8	6.1	5.1	8.8	5.1	5.1	7.0	6.8	4.2	7.4	6.1	4.1	5.9	4.9
Informing, involving and listening to children	7.2	9.2	7.7	6.3	9.1	6.8	6.4	8.3	7.8	5.4	8.1	7.1	6.0	6.9	7.5
Clear commitment to Signs of Safety from management	7.6	8.9	7.8	6.9	8.8	7.9	6.5	8.3	8.4	5.3	7.8	6.4	6.0	7.0	7.7
It feels like a safe organisation in which to work	7.2	8.7	7.0	7.4	8.4	7.8	6.7	8.3	8.8	5.3	7.6	6.1	5.2	6.2	7.8
Family Court engaged in Sigs of Safety approach	4.1	7.2	5.3	4.7	6.8	4.2	2.2	5.2	6.8	2.9	5.9	3.0	1.6	4.9	2.9
Recording systems aligned with Signs of Safety	6.5	8.2	7.7	5.3	7.6	5.8	4.9	8.5	7.9	3.5	6.6	4.6	4.4	6.0	4.1
Quality assurance systems aligned with Signs of Safety	5.9	8.3	6.4	5.3	8.1	5.1	5.7	7.4	8.0	4.1	7.3	5.7	4.0	5.2	6.6

Elements of SofS	P9	P9	P9	P7	P7	P7	P8	P8	P8	P2	P2	P2	P1	P1	P1
Group supervision and appreciative inquiry in place	4.4	7.6	6.3	4.5	7.4	4.6	4.5	7.1	7.0	4.9	7.8	7.3	3.3	5.6	4.3
Partner agency involvement	4.4	7.5	6.7	4.7	7.6	5.5	4.9	7.6	6.5	4.3	7.3	5.6	4.7	6.2	2.8

Appendix 9: Outcomes analysis

Research objectives

The aim of the analysis is to assess whether being an SofS pilot site has had a measurable impact on outcomes for children compared to sites that are not using SofS. For each outcome we are examining:

1. Is there a difference in outcome between the pilot and comparison (both in any particular year and across all years)?
2. Does the outcome change over time?
3. Is the change in outcome over time different between the pilot and comparison sites?
4. Are there patterns in the category of any outcomes that do appear different between pilot and comparison sites?

Methods

Selecting comparison groups

In phase 1 the 10 pilot sites were compared to 10 authorities made up of the closest statistical nearest neighbour[2] to each pilot site. In phase 2 we wanted to refine this process by ensuring that none of the authorities in the comparator sites were reported as using SofS. The NIHR Health and Social Care Workforce Research Unit (HSCWRU) surveyed LAs in 2018 as to their use of SofS and we excluded LAs which identified themselves as using 'pure' SofS or 'some form' of SofS. These data are incomplete and are also likely to be out of date for some areas, and so desk research was also undertaken in January 2020 to ensure that none of the selected comparison sites mentioned SofS in their latest children plans.

In 2019 1 of the 10 pilot sites dropped out, leaving only 9 sites. In 8 of these sites, the closest nearest neighbours reported not to be using SofS were selected. In the remaining site there were no nearest neighbours that did not use SofS and so we decided to exclude this site from the outcomes analysis due to lack of suitable comparator. A breakdown of the pilot and comparator sites by broad region, type and CSC grade is provided in Table A9.1.

[2] Based on the Children's Services Statistical Neighbour Model available at www.gov.uk/government/publications/local-authority-interactive-tool-lait

It should be noted that one of the sites only became a ‘pilot’ at the beginning of EIP2, although they had been using SofS prior to becoming a pilot. All other sites included in the outcomes analysis were pilots during EIP1.

Table A9.1: Characteristics of pilot and SNN authorities

Authority type/Region/Ofsted rating	Pilot sites	SNNs
English unitary	2	3
London borough	2	1
Shire county	4	4
East/South East	4	4
London	2	1
East/West Midlands	1	1
South West	1	2
Inadequate	1	2
Requires improvement	4	2
Good	1	3
Outstanding	2	1

Please note that this includes only LAs included in the quantitative analysis and therefore excludes one of the pilot sites due to lack of statistical nearest neighbour.

When calculating statistical nearest neighbours, each LA is measured in terms of its likeness to every other. The distance between any 2 LAs is defined as the weighted Euclidean distance between the authorities using each of the background variables. All but one of the pilot sites had a suitable SNN that was ‘very close’,[3] with the remaining site being ‘close’ to its SNN.

Identifying outcome data

In 4 of the outcomes categories (Reducing risk for children, Reducing days spent in state care, Increasing workforce wellbeing, Increasing workforce stability) the relevant outcome measures were identified from the provisional shared indicator list for evaluating Innovation Programme and PiP projects. Data from 2014/15, 2015/16, 2016/17, 2017/18 and 2018/19 were collated and re-coded to allow for relevant analysis. This analysis was purposely broad initially to identify possible trends in the data to provide focus for the DiD analysis by What Works for Children’s Social Care (WWCSC).

[3] Defined as weighted Euclidean distance between local authorities being equivalent to less than 0.55 per standardised variable

We have not examined outcomes under ‘Increasing wellbeing for children and families’ or ‘Creating greater stability for children’ due to a lack of available consistent data over the relevant 5-year period.

For the outcome category ‘Generating better value for money’, data on the costs involved in implementation and maintaining SofS were collected from each of the pilots. This was compared with any potential savings identified by applying unit costs to any significant outcome impacts that were observed in the pilots.

To avoid placing any additional burden on LAs we sourced all of the secondary data from existing national collections. The indicator list was checked against both the MTM Core Dataset and the provisional shared indicator list for evaluating IP and PiP projects.

The final analysis included 29 key outcome indicators which have been grouped according to the 7 outcome categories identified by the DfE as a particular focus for the current evaluations. Table A9.2 provides an overview of the number of outcomes included in each category.

Table A9.2: Category and number of indicators

Categories	Outcome indicators
Reduce risk for CYP: Children in need	3
Reduce risk for CYP: Referrals and re-referrals	8
Reduce risk for CYP: Child protection plans	7
Reduce days spent in state care	5
Increase staff wellbeing	3
Reduce staff turnover and agency rates	3

Analysis methodology

The following analysis approach was undertaken for each of the 29 outcomes.

- Descriptive statistics and confidence intervals were produced for the pilot and comparison groups. These were plotted onto a graph to examine possible patterns or trends.
- To test whether there was a significant difference in outcome in any one year between pilot and comparison groups, a one-way analysis of variance (ANOVA) was used, with group as the factor variable and outcome for each year individually as the dependent variable.
- To test whether there is significant change over time and also whether there is a difference between pilot and comparison group across time, a 2-way analysis of

variance (ANOVA) was used, with outcome as the dependent variable and group and year as factor variables (having converted data from wide to long format).

- Where there was a significant difference, post-hoc testing was used to establish an estimate of the effect size.

Changes to evaluation methods

We have removed 3 outcomes from the 'Increasing wellbeing for children and families' group for the final analysis as the data did not cover the full period and initial analysis had not suggested any significant impacts in this area.

We added an outcome examining special guardianship to ensure the analysis aligned as closely as possible with the WWCCSC additional DiD analysis.

We removed one of the pilot sites from the analysis due to lack of a suitable comparator.

Key challenges

Data for some of the indicators were not available across all relevant years due to changes in data collections and/or timing of publication.

Identifying comparison sites in which SofS was never used was complicated by high staff movement between LAs (which resulted in some staff in non-SofS sites using SofS methods).

Outcomes

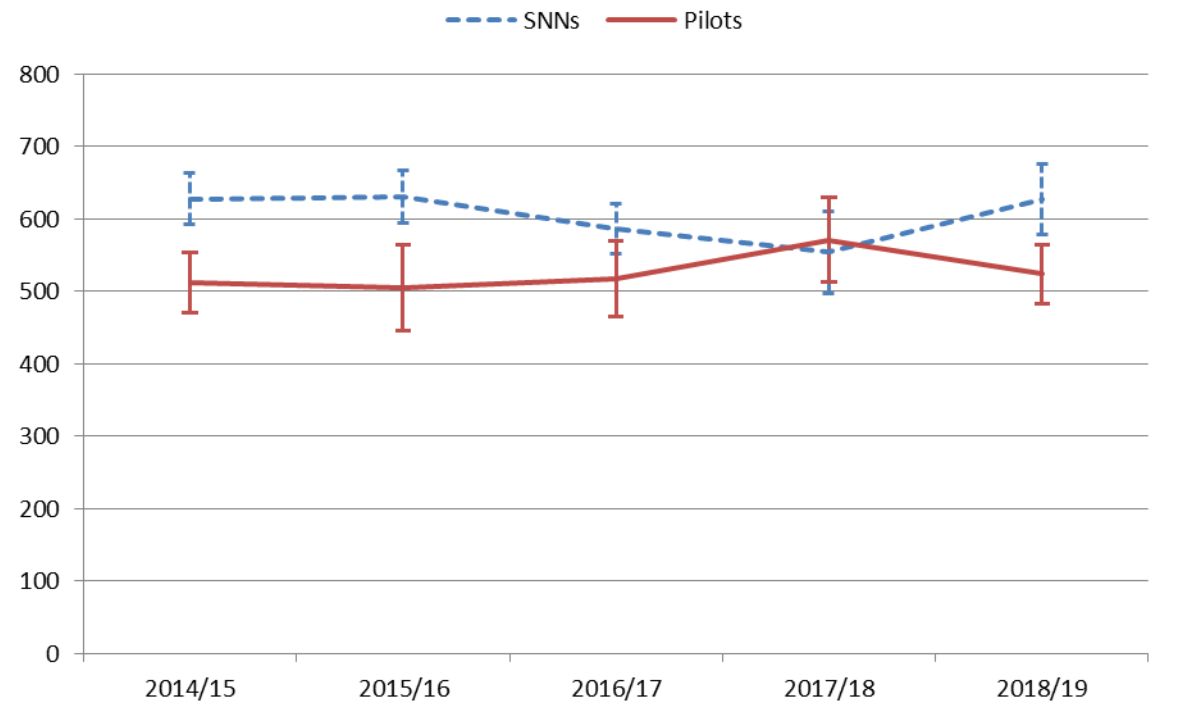
Reduce risk for children and young people: Children in need

This category contains 3 outcomes indicators:

- Children in need throughout the year (rate per 10,000 children)
- Children in need – case duration less than 3 months (%)
- Children in need – case duration more than 2 years (%)

Children in need throughout the year (rate per 10,000 children) was significantly lower in pilots than the SNNs across all years (p=.008), with pilot status a significant effect (p=.01) in the 2-factor analysis. However, this effect did not vary significantly over time (that is, there was no interaction effect with year), as can be seen in Figure A9.1.

Figure A9.1: Children in need throughout the year (rate per 10,000 children)



Reduce risk for children and young people: Referrals and re-referrals

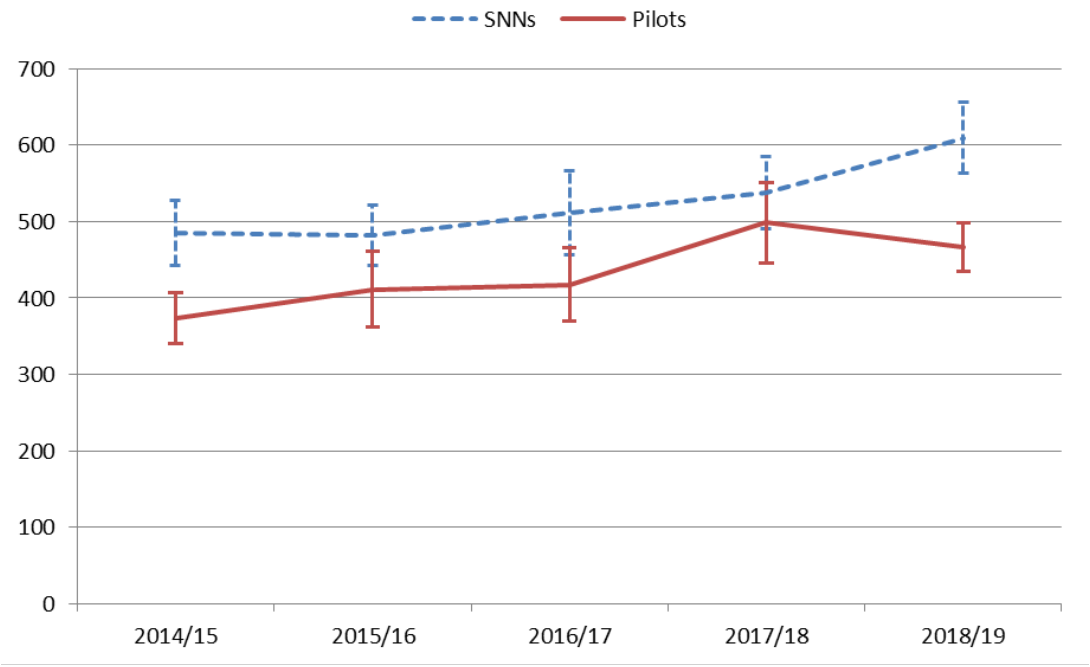
This category contains 8 outcomes indicators:

- Referrals (rate per 10,000 children)
- Referrals within 12 months of previous referral (%)
- Referrals resulted in no further action (%)
- Referrals where the child was assessed not to be in need (%)
- Assessments (rate per 10,000 children)
- Median duration of assessment (working days)
- Assessments that started and finished on same day (%)
- Assessments that lasted 61 days or more (%)

There were no significant differences between the pilots and SNNs in any of the 4 outcome measures related to referrals.

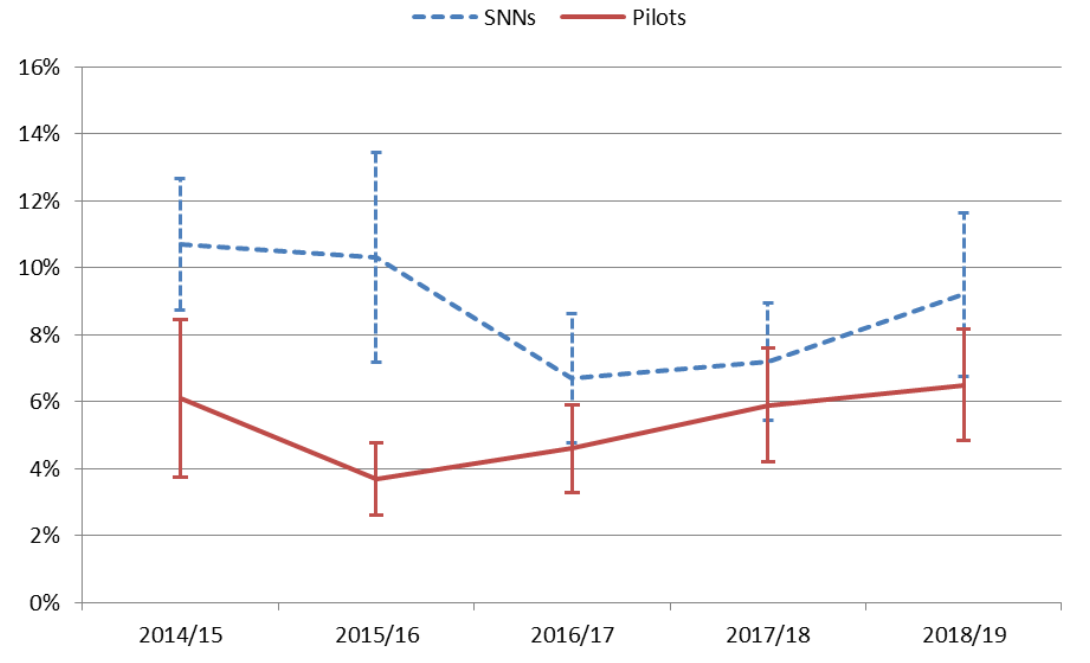
Assessments rate per 10,000 children was significantly lower in pilots than the SNNs across all years (p=.002), with pilot status a significant effect (p=.002) in the 2-factor analysis. This effect did not vary significantly over time as there was a marked increase in the rate in the past year in both SNNs and SofS sites, despite a fall in the latest year in SofS sites.

Figure A9.2: Assessment rate per 10,000 children



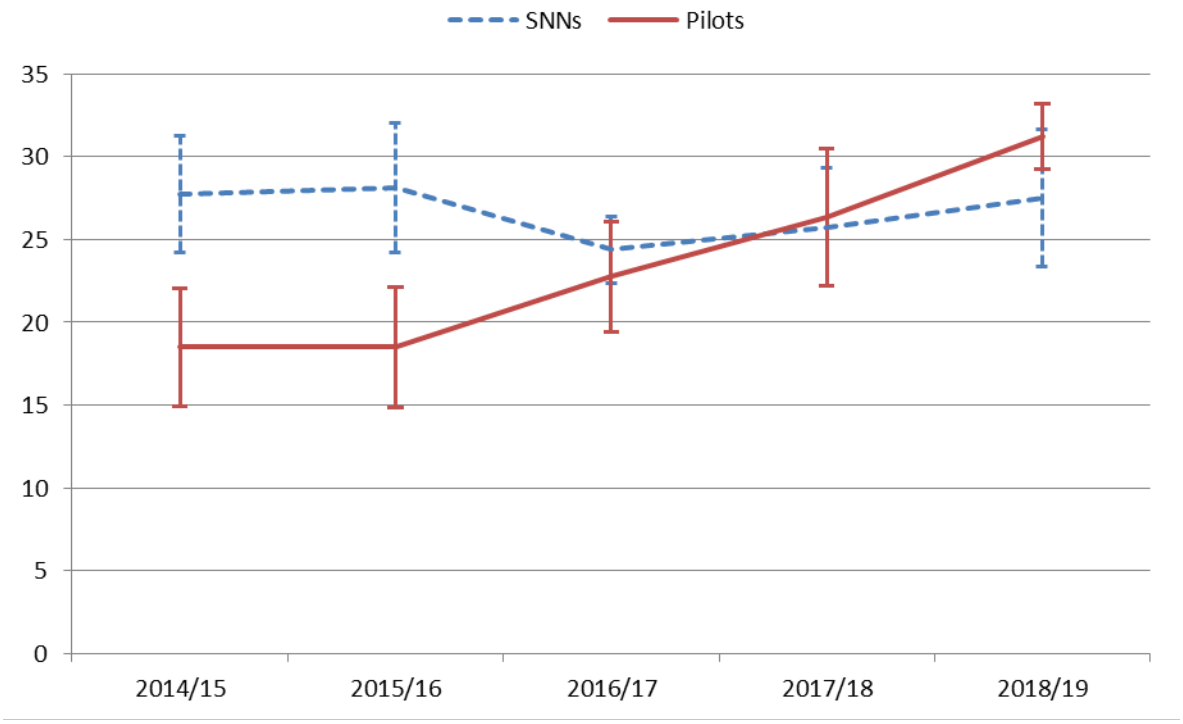
The percentage of assessments that lasted 61 days or more was significantly lower in pilots than the SNNs across all years (p=.007), with pilot status a significant effect (p=.008) in the 2-factor analysis. This effect did not vary significantly over time despite a reduction in the percentage within SNNs, as shown in Figure A9.3.

Figure A9.3: Assessments that lasted 61 days or more (%)



Median duration of assessments was not significantly different across the years in pilot sites (p=.152), nor did it change significantly in the 2-factor analysis.

Figure A9.4: Median duration of assessments



Pilot sites no longer had shorter duration of assessments and the difference in those assessments lasting 61+ days had reduced in the past few years. The overall rate of assessments in pilot sites was lower, but this had been true in each of the past 5 years.

Reduce risk for children and young people: Child protection plans

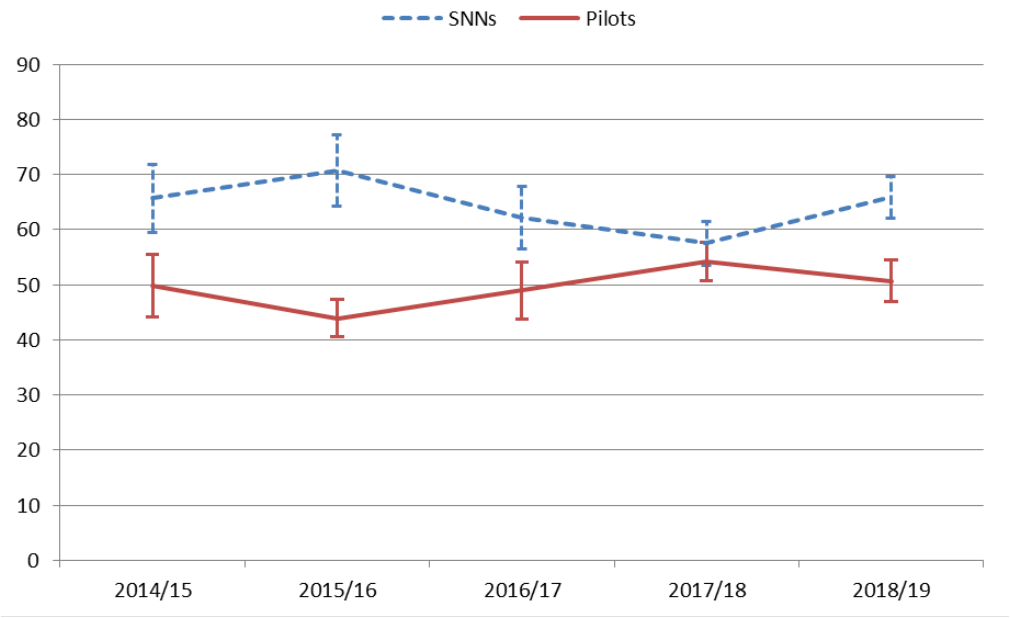
This category contains 7 outcome indicators:

- Section 47 enquiries throughout the year (rate per 10,000 children)
- Child protection (CP) conferences throughout the year (rate per 10,000 children)
- Duration between the start of section 47 enquiries and initial CP conference (working days)
- Child protection plans at March (rate per 10,000 children)
- Children who became the subject of a plan for a second or subsequent time (%)
- Child protection plans at March with case duration of 3 months or less (%)
- Child protection cases reviewed within the required timescales (%)

Child protection conferences throughout the year (rate per 10,000 children) are significantly lower in pilots than the SNNs across all years ($p < .001$), with pilot status a significant effect ($p < .001$) in the 2-factor analysis.

There was no significant change across years and the interaction between year and SofS use was not significant, although there was a noticeable trend which saw the rate within pilots fall in 2015/16 and then increase in the following years until it fell in 2018/19 while the rate in SNNs was increased slightly in 2015/16 before falling in the following years and increasing in 2018/19, as shown in Figure A9.5. This resulted in the child protection conference rate in the pilots being significantly lower (p=.01) in 2018/19 than in the SNNs.

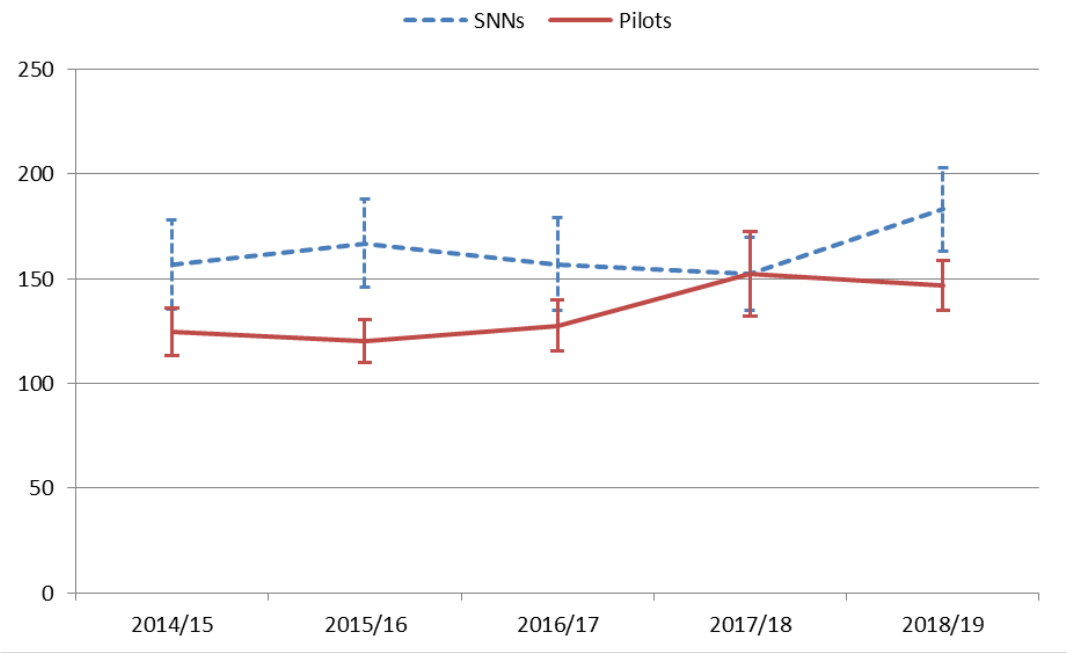
Figure A9.5: Child protection conferences throughout the year (rate per 10,000 children)



A near identical trend was also observed in the rate of child protection plans, which were also significantly lower in pilots than the SNNs across all years ($p < .001$), with pilot status a significant effect ($p < .001$) in the 2-factor analysis.

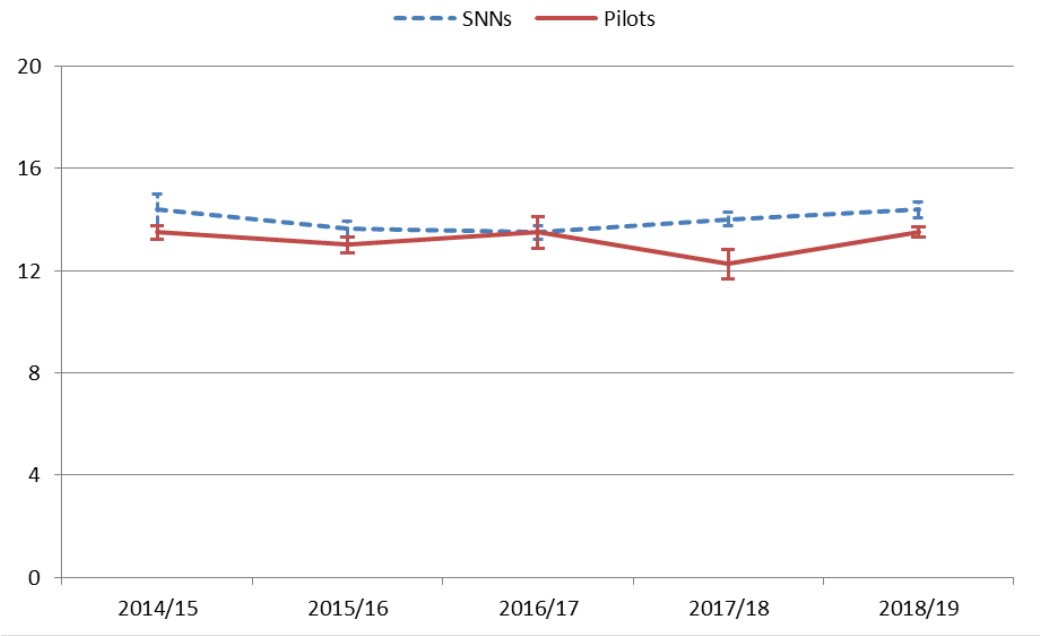
Section 47 enquiries throughout the year (rate per 10,000 children) were significantly lower in pilots than the SNNs across all years (p=.009), with pilot status a significant effect (p=.011) in the 2-factor analysis. The effect did not vary significantly over time despite recent changes in the rate in both groups (in different years) as seen in Figure A9.6.

Figure A9.6: Section 47 enquiries throughout the year (rate per 10,000 children)



Duration between the start of section 47 enquiries and initial CP conference (median working days) was significantly lower in pilots than the SNNs across all years (p=.002), with pilot status a significant effect (p=.002) in the 2-factor analysis. This effect did not vary significantly over time, although the rate in pilot sites decreased markedly in 2017/18, as shown in Figure A9.7.

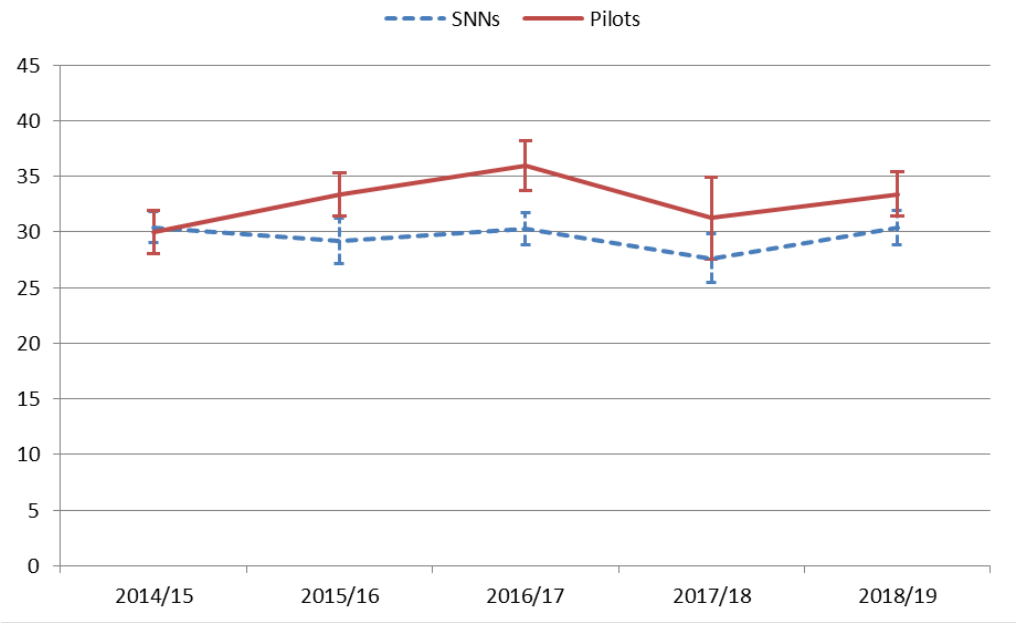
Figure A9.7: Duration between start of section 47 enquiries and ICPCs (median working days)



Child protection plans at March with case duration of 3 months or less (%) were significantly (p=.017) higher for the SofS pilots with main effect of pilot status also

significant (p=.019) in the 2-factor analysis. This effect did not vary significantly over time despite the 2 groups having a similar percentage in 2014/15 which stayed relatively stable in SNNs while in SofS sites it increased in the 2 years to 2016/16 before decreasing in 2017/18.

Figure A9.8: Child protection plans at 31 March with case duration of 3 months or less (%)



Pilot sites had lower rates of child protection conferences and child protection plans than their SNNs, although this was the case in each of the previous 5 years. Child protection plans in pilot sites were also more likely to have a case duration of 3 months or less, although this difference reduced slightly towards the end of the project.

Reduce days spent in state care

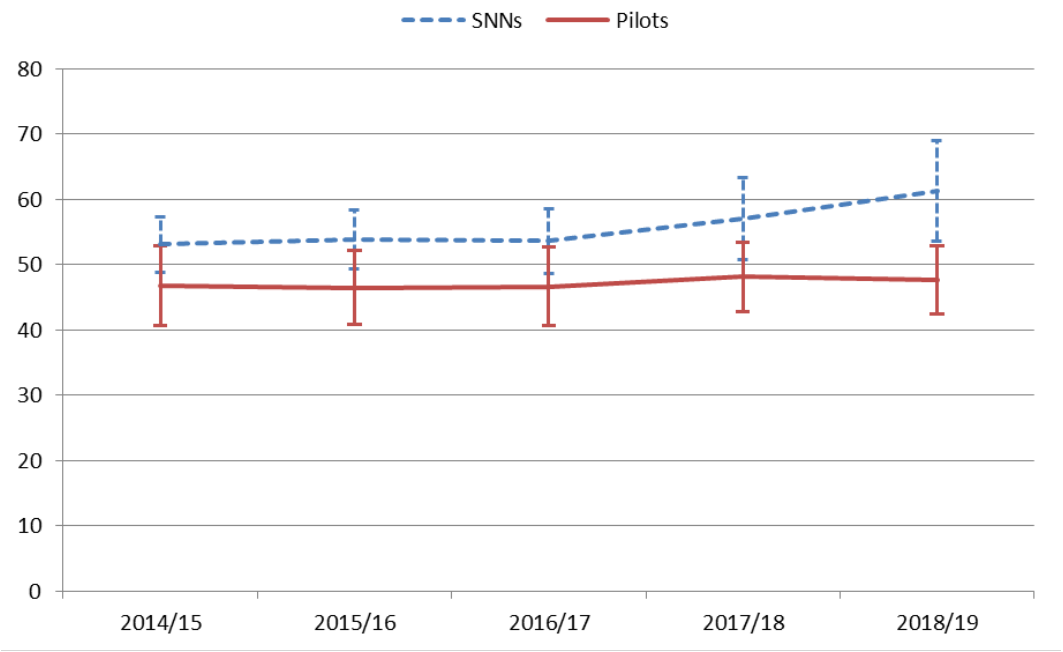
This category contains 5 outcome indicators:

- CAFCASS care application demands (rate per 10,000 children)
- Looked after children at March (rate per 10,000 children)
- Number of children becoming looked after (rate per 10,000 children)
- Percentage of looked after children adopted during the year
- Children who ceased to be looked after – percentage of special guardianship orders

There were no significant differences between the pilots and SNNs in care application demands, the number of children becoming looked after, the percentage of looked after children adopted during the year or the percentage of children who ceased to be looked after through special guardianship orders.

The looked after children rate was significantly lower in pilots than the SNNs across all years (p=.01) and in the 2-factor analysis with pilot status (p=0.019) and year. There was no significant change across years and the interaction between year and SofS use was not significant.

Figure A9.9: Looked after children at 31 March (rate per 10,000 children)



There is no evidence in the outcomes for looked after children that SofS use has any impact on the number of days children spend in care.

Increase staff wellbeing

This category contains 3 outcome indicators:

- Number of children in need per children's social worker
- Sickness absence rate (%)
- Caseload: average cases per social worker

Caseload data was not available for 2014/15 and 2015/16.

There were no significant differences between pilots and SNNs in any of the 3 measures in any individual year, nor across the years as a whole, with no significant interaction effect. As such there is no evidence that SofS use had any impact on caseload or staff sickness levels.

Reduce staff turnover and agency rates

This category contains 3 outcome indicators:

- Children's social workers - vacancy rate (%)
- Children's social workers - turnover rate (%)
- Children's social workers – agency worker rate (%)

There were no significant differences between pilots and SNNs in any of the 3 measures in any individual year, nor across the years as a whole with no significant interaction effect. As such there is no evidence that SofS use had any impact on the retention of staff nor the use of agency staff.

Summary tables

Table A9.3 presents a comparison of average outcome measurements in year 2018/19 (September 2019 for workforce outcomes) between the SofS pilot sites and their SNNs. F statistics are computed using one-way analysis of variance (ANOVA) with pilot status as the fixed factor.

Table A9.4 presents the comparison of averages of the outcome variables for the SofS (pilot councils) and SNN authorities across all years. F-statistics and p-values were obtained from 1-way analysis of variance as before with pilot status as the independent variable.

Table A9.5 presents the ANOVA tables for 2-factor analysis with pilot status and year (2014/15–2018/19) as 2 factors. It also includes an interaction effect to assess if the effect of the authorities is different over the years or not.

Table A9.3: Comparison of outcome variables for 2018/19 between SofS and SNN authorities

Outcomes	Pilot mean	SNN mean	F Statistic	P-value
Children in need throughout the year (rate per 10,000 children)	523.7	627.4	2.7	0.12
Children in need - case duration less than 3 months (%)	29.4	26.0	1.8	0.20
Children in need - case duration more than 2 years (%)	29.4	30.9	0.3	0.57
Referrals (rate per 10,000 children)	480.3	543.9	1.2	0.29
Referrals where the child was assessed not to be in need (%)	33.6	25.2	2.0	0.18
Referrals resulted in no further action (%)	5.4	8.4	0.6	0.44

Outcomes	Pilot mean	SNN mean	F Statistic	P-value
Referrals within 12 months of previous referral (%)	23.0	22.7	0.0	0.92
Assessments rate per 10,000 children	466.4	609.7	6.5	0.02
Median duration of assessment (working days)	31.3	27.5	0.7	0.43
Assessments that started and finished on same day (%)	2.0	2.8	0.6	0.44
Assessments that lasted 61 days or more (%)	6.5	9.2	0.8	0.38
Section 47 throughout the year (rate per 10,000 children)	146.9	183.0	2.4	0.14
*Child protection conferences throughout the year (rate per 10,000 child)	50.7	65.9	7.9	0.01
Duration between start of section 47 enquiries and initial CP conference (days)	13.5	14.4	5.4	0.04
*Child protection plans at March (rate per 10,000 children)	43.6	57.6	9.1	0.01
Children who became the subject of a plan for a subsequent time (%)	20.6	19.4	0.5	0.49
Child protection plans at March with case duration of 3 months or less (%)	33.4	30.4	1.5	0.25
Child protection cases reviewed within the required timescales (%)	93.0	93.1	0.0	0.97
Number of children in need per children's social worker 2019	12.2	14.1	1.9	0.19
Children's social workers – turnover rate (%) 2019	13.8	18.6	1.8	0.20
Children's social workers – vacancy rate (%) 2019	17.6	16.5	0.0	0.84
Children's social workers – agency worker rate (%) 2019	17.9	17.0	0.0	0.87
Children's social workers – sickness absence rate 2019	2.7	2.8	0.0	0.96
Caseloads: Number of cases per social worker 2019	16.8	17.1	0.2	0.69
Looked after children at March (rate per 10,000 children)	47.6	61.3	2.1	0.17

Outcomes	Pilot mean	SNN mean	F Statistic	P-value
Number of CYP becoming looked after per 10,000 children in LA	20.1	25.8	1.8	0.20
Percentage of looked after children adopted during the year	11.7	11.7	0.0	1.00
Care applications per 10,000 child population	8.4	10.3	1.9	0.19
Children who ceased to be looked after – special guardianship order %	12.4	11.7	0.1	0.79

Variables marked with * remain significant when Pilots 1 and 2 and their SNNs are removed from the analysis.

Table A9.4: Comparison of outcome variable across all years between SofS and SNN authorities

Outcomes	Pilot mean	SNN mean	F Statistic	P-value
Children in need throughout the year (rate per 10,000 children)	525.7	605.1	7.5	0.01
Children in need - case duration less than 3 months (%)	28.0	26.6	1.4	0.24
*Children in need - case duration more than 2 years (%)	28.5	30.8	2.9	0.09
Referrals (rate per 10,000 children)	473.4	516.7	2.5	0.12
Referrals where the child was assessed not to be in need (%)	28.7	24.9	1.5	0.22
Referrals resulted in no further action (%)	8.6	9.0	0.0	0.84
Referrals within 12 months of previous referral (%)	21.5	22.8	1.2	0.28
Assessments rate per 10,000 children	433.4	525.3	10.1	0.00
Median duration of assessment (working days)	23.5	26.7	2.1	0.15
Assessments that started and finished on same day (%)	5.1	3.1	3.2	0.08
Assessments that lasted 61 days or more (%)	5.4	8.8	7.7	0.01
*Section 47 throughout the year (rate per 10,000 children)	134.3	163.2	7.2	0.01
*Child protection conferences throughout the year (rate per 10,000 child)	49.5	64.4	23.6	0.00
*Duration between start of section 47 enquiries and initial CP conference (days)	13.2	14.0	9.8	0.00
*Child protection plans at March (rate per 10,000 children)	42.7	55.9	26.1	0.00

Outcomes	Pilot mean	SNN mean	F Statistic	P-value
Children who became the subject of a plan for a subsequent time (%)	19.5	19.5	0.0	0.99
Child protection plans at March with case duration of 3 months or less (%)	32.8	29.6	5.9	0.02
Child protection cases reviewed within the required timescales (%)	94.3	93.2	0.8	0.37
Number of children in need per children's social worker 2019	13.8	14.2	0.3	0.58
Children's social workers – turnover rate (%) 2019	17.3	17.4	0.0	0.98
Children's social workers – vacancy rate (%) 2019	19.0	17.6	0.4	0.52
Children's social workers – agency worker rate (%) 2019	17.6	16.7	0.1	0.71
Children's social workers – sickness absence rate 2019	3.0	3.1	0.1	0.73
Caseloads: Number of cases per social worker 2019	16.7	17.1	0.2	0.64
Looked after children at March (rate per 10,000 children)	47.1	55.8	6.3	0.01
Number of CYP becoming looked after per 10,000 children in LA	22.5	25.4	2.8	0.10
Percentage of looked after children adopted during the year	13.4	14.4	0.8	0.39
Care applications per 10,000 child population	9.1	10.0	1.4	0.25
Children who ceased to be looked after – special guardianship order (%)	12.1	11.9	0.1	0.82

Variables marked with * remain significant when Pilots 1 and 2 and their SNNs are removed from the analysis.

Table A9.5: Significant ANOVA results for outcome variables comparing pilot status across years

Dependent variable: Children in need throughout the year (rate per 10,000 children)

Source	Type III sum of squares	df	Mean square	F	Sig.
Year	5225.143	4	1306.286	.073	.990
Pilot status	125960.192	1	125960.192	7.054	.010
Year * pilot status	53780.322	4	13445.080	.753	.559
Error	1250035.983	70	17857.657		
Corrected total	1435001.640	79			

Dependent variable: Assessments rate per 10,000 children

Source	Type III sum of squares	df	Mean square	F	Sig.
Year	140012.864	4	35003.216	2.145	.084
Pilot status	168930.580	1	168930.580	10.354	.002
Year * pilot status	24801.581	4	6200.395	.380	.822
Error	1142063.892	70	16315.198		
Corrected total	1475808.918	79			

Dependent variable: Assessments that lasted 61 days or more (%)

Source	Type III sum of squares	df	Mean square	F	Sig.
Year	.008	4	.002	.590	.671
Pilot status	.024	1	.024	7.365	.008
Year * pilot status	.007	4	.002	.562	.691
Error	.227	70	.003		
Corrected total	.266	79			

Dependent variable: Section 47 throughout the year (rate per 10,000 children)

Source	Type III sum of squares	df	Mean square	F	Sig.
Year	6574.288	4	1643.572	.679	.609
Pilot status	16669.538	1	16669.538	6.890	.011
Year * pilot status	4925.691	4	1231.423	.509	.729
Error	169355.083	70	2419.358		
Corrected total	197524.599	79			

*Dependent variable: Child protection conferences throughout the year (rate per 10,000 children)

Source	Type III sum of squares	df	Mean square	F	Sig.
Year	93.199	4	23.300	.122	.974
Pilot status	4429.776	1	4429.776	23.121	.000
Year * pilot status	1111.903	4	277.976	1.451	.227
Error	13411.361	70	191.591		
Corrected total	19046.240	79			

*Dependent variable: Duration between start of section 47 enquiries and initial CP conference (median working days)

Source	Type III sum of squares	df	Mean square	F	Sig.
Year	8.625	4	2.156	1.621	.179
Pilot status	13.613	1	13.613	10.232	.002
Year * pilot status	6.325	4	1.581	1.189	.323
Error	93.125	70	1.330		
Corrected total	121.688	79			

*Dependent variable: Child protection plans at March (rate per 10,000 children)

Source	Type III sum of squares	df	Mean square	F	Sig.
Year	75.877	4	18.969	.136	.968
Pilot status	3523.185	1	3523.185	25.347	.000
Year * pilot status	719.302	4	179.825	1.294	.281
Error	9729.726	70	138.996		
Corrected total	14048.090	79			

*Dependent variable: Child protection plans at March with case duration of 3 months or less (%)

Source	Type III sum of squares	df	Mean square	F	Sig.
Year	132.043	4	33.011	.911	.462
Pilot status	208.013	1	208.013	5.741	.019
Year * pilot status	82.201	4	20.550	.567	.687
Error	2536.243	70	36.232		
Corrected total	2958.500	79			

Dependent variable: Looked after children at March (rate per 10,000 children)

Source	Type III sum of squares	df	Mean square	F	Sig.
Year	252.800	4	63.200	.245	.912
Pilot status	1496.450	1	1496.450	5.791	.019
Year * pilot status	137.300	4	34.325	.133	.970
Error	18089.250	70	258.418		
Corrected total	19975.800	79			

Variables marked with * remain significant when Pilots 1 and 2 and their SNNs are removed from the analysis.

Appendix 10: Difference-in-differences analysis

Methods

We used a DiD analysis which compares trends in children's social care outcomes in the pilot LAs with trends in outcomes in the comparator LAs. If the trends in outcomes were similar before the treatment is introduced, we would expect the parallel trends to continue. Where we see a change in the trends in the treatment group or control group but not the other, we can interpret this as a causal estimate of the treatment effect on the treated, in other words, the change can be attributed to the implementation of SofS. DiD is a common technique to evaluate social interventions which have already commenced (as was the case here) and / or where randomisation is not deemed to be appropriate. We matched the pilot LAs to up 2 comparator LAs that most closely resembled them in terms of their trends in the respective outcomes before the introduction of SofS. Within these pilot and comparator pairs, we then matched individuals in pilot and comparator pairs using coarsened exact matching (CEM).[4] We ran our analysis based on this matched individual-level dataset taking into account both individual-level and local authority level characteristics. We evaluated the impact of SofS on 4 outcomes, 1 of which was identified as showing promise in the Round 1 evaluation and the other three of which were based on consultation with MTM. We refer to these outcomes as EQ3a (duration of assessment), EQ3b (re-referrals), EQ3v (re-referrals that lead to CPP/LAC) and EQ3d (kinship care instead of non-kinship care).

Regression specification

We estimate the effect of the intervention, β_3 on the outcomes of interest Y_ist:

$$Y\_ist = \alpha + \beta\_1\, SS\_s + \beta\_2\, T\_st + \beta\_3\, (SS\_s \cdot T\_st\,) + \beta\_4\, X\_ist + \beta\_5\, \Gamma\_st + \ [\![\theta Year]\!]\ \_t\ \ [\![+ \epsilon]\!]\ \_ist \qquad [1]$$

Where:

- α is a regression constant
- SS_s is a binary indicator of whether the local authority received SofS
- T_st is a binary indicator of whether the local authority is receiving SofS at time t
- X_ist is a vector of participant level characteristics
- Γ_st is a vector of local authority level characteristics

[4] This matching method involves temporarily coarsening each control variable.

- $[\![Year]\!]\_t$ are year dummy variables to capture time trends common to all local authorities (these are excluded where they are perfectly collinear with T_t)
- ϵ_ist is an error term, denoting standard errors clustered at the level of the local authority (the level at which assignment takes place).

We run linear regressions except for the evaluation questions with small proportions of true outcomes (<5%), where we run logistic regressions as our main analysis (this is the case for EQ3c).

Defining pre- and post-treatment

We compared data from before local authorities started implementing SofS, and from after what we call the 'settling-in period'. The settling-in period describes the time after local authorities started implementing SofS, where local authorities and their teams understand and become accustomed to new ways of working with the model. During this period, we expect changes observed in the outcome to reflect the period of change rather than the impact of SofS 'business as usual'. Including this period could overestimate the impact of SofS, for example, by picking up initial enthusiasm which then tails off, or including this period could underestimate the impact of SofS, as staff are still familiarising themselves with the new way of working. We were interested in picking up the impact of SofS in its 'business as usual' guise, which does not necessarily mean 'perfect implementation'. Estimating the settling-in period comes from triangulation of qualitative work conducted by Dr Mary Baginsky, a survey of the pilot sites at a SofS leadership day, individual follow-ups with the pilot local authorities and an assessment of what is reasonable. At the start of the Innovation Programme, it was anticipated that a 2-year period of intense activity within the context of a longer-term commitment, estimated to be about 5 years, would be needed to embed SofS. The lengths of the settling in period varied from nearly 2 to 5 years.

Sensitivity analysis

We ran several sensitivity analyses to assess the robustness of our main results.Since a DiD analysis rests on the assumption that the pilot and comparator LAs had parallel trends in outcomes prior to the implementation of SofS, we conducted a placebo test. We tested for a 'treatment effect' prior to the start of the intervention by including leads of the DiD indicator. Where the majority of coefficients of the leads of the DiD indicator were insignificant, we took that as evidence of parallel trends. The parallel trends test holds for evaluation questions a, b and d.

To assess whether the DiD results are sensitive to the timing of the settling-in period chosen or whether the impact of SofS changes over time, we also introduce lagged

treatment variables: where the lagged DiD indicators are significant, this suggests SofS has an impact on the outcome but later than we anticipated.

Secondary analysis

In addition to evaluating whether SofS 'works' overall in the primary analysis, another question of interest is whether it works in some circumstances and not others, or particularly well in some circumstances. We investigated whether SofS had a differential impact on LAs by Ofsted rating where we categorised the LA as having a 'high' Ofsted rating in a particular year if their most recent Ofsted rating for children's services overall was Good or Outstanding and as having a 'low' Ofsted rating in a particular year if their most recent Ofsted rating for children's services overall was Requires Improvement or Inadequate. We run triple difference analysis interacting the Ofsted ratings of the local authorities with the standard DiD indicator.

We also investigated whether the treatment effect is stronger for pilot sites that have achieved a better delivery of the model, as measured by self-reported profiling scores of the embeddedness of SofS, and MTM's scoring of the quality of delivery of SofS. The analysis was conducted by scaling the SofS variable from 0 (not delivered) to 1 (delivered the most well), scaled by the highest score. If the coefficient on the standard DiD indicator in the main analysis is insignificant but the coefficient on the interaction of self-reported embeddedness or MTM's delivery score with T_st is significant, we interpret this as SofS having a significant impact when better embedded or delivered as appropriate, the idea being that one would expect SofS in a higher 'dose' to have a larger effect.

We recognise that there may be overlap between the SofS model and at least 2 other popular practice models, Restorative Practice and Reclaiming Social Work. If the mechanisms through which the 3 practice models are expected to affect outcomes are the same, comparing the pilot sites to local authorities using Restorative Practice or Reclaiming Social Work would bias downwards any potential effect of SofS. We ran additional analysis to investigate this possibility, where SS_s, the binary indicator for participation in SofS in the main analysis ([1]) becomes a 3-level indicator of whether the local authority received SofS (the base category), a similar practice model (Restorative Practice and Reclaiming Social Work), or neither:

$$Y\_ist = \alpha + \beta\_1 \ [\![Alernative\ PM]\!]\_s + [\![\beta\_2 \ [\![Neither\ PM]\!]\_s + \beta]\!]\_3\, T\_st + [\![\beta\_4\, ([\![Alternative\ PM]\!]\_s \cdot T\_st) + \beta]\!]\_5\, ([\![Neither\ PM]\!]\_s \cdot T\_st) + \beta\_6\, X\_ist + \beta\_7\, \Gamma\_st + [\![\theta Year]\!]\_t \ [\![+ \epsilon]\!]\_ist \quad [5]$$

We interpret the negative of a significant coefficient on $[\![Neither\ PM]\!]\_s \cdot T\_st$, β_5, as the impact of SofS on the outcome of interest compared to comparator LAs without a similar practice model. It is the negative because the base category is SofS, the reverse

of in the main analysis. Since we did not have access to the exact implementation dates of the alternative practice models for each local authority, the T_st of $〚Alternative\ PM〛\_s \cdot T\_st$ relates to whether the local authority was receiving SofS at time t rather than the practice model as we did not have information on when the comparators started using the alternative practice model and so the results for this analysis are indicative only.

Main analysis

Matching results

Upon conclusion of the local authority level matching, we used the following pilots for our main analysis. The number of pilots used for each evaluation question varies considerably. This is the result of the local authority level matching on pre-SofS trends of the outcome variable (or a proxy thereof).

Table A10.1: Key to pilots' involvement in each evaluation question assessed through the DiD analysis

Pilots	EQ3a	EQ3b	EQ3c	EQ3d
Pilot 1	x	✓	x	x
Pilot 2	✓	✓	✓	✓
Pilot 3	✓	✓	✓	✓
Pilot 4	✓	x	x	x
Pilot 5	✓	x	x	✓
Pilot 6	x	✓	x	x
Pilot 7	✓	x	x	✓
Pilot 8	x	x	x	✓
Pilot 9	x	x	x	✓

Within the groups of matched LAs, we then conducted coarsened exact matching to provide a more efficient estimate of the impact of SofS. We assessed the quality of the resulting balance using the multivariate imbalance scores (where 0 represents a perfectly balanced sample) and the local common support (where 100% represents that all individuals in pilots have a counterpart with the exact same demographics in a

comparator LA). The resulting multivariate imbalance scores for the different populations are reported in Table A10: 2 below.

The scores are calculated using the uncoarsened dataset. After matching, the imbalance score of the coarsened dataset would be equal to zero. The matching improves the balance of the data but does not fully account for the differences between the pilot and comparator LAs so we also control for individual-level covariates in the regressions.

Table A10.2: Multivariate imbalance scores and local common support pre and post CEM

EQ	L1 imbalance score		Local common support	
	Prior to CEM	Post CEM	Prior to CEM	Post CEM
1	0.62	0.61	48%	64%
2	0.39	0.37	41%	56%
3	0.36	0.34	41%	56%
4	0.55	0.47	29%	48%

Data source: Office for National Statistics – National Pupil Database. The L1 imbalance score and the local common support are calculated for the original population prior to matching, and for the matched population after conclusion of the CEM for the uncoarsened variables. Note that the imbalance score is 0 when it is calculated using the coarsened variables.

Summary statistics

Since we use different populations and different pilot LA comparator LA matches for the 4 evaluation questions, the sample size and covariate balance varies between the relevant samples. Table A10: 3 below gives an overview of each individual sample. Means are only reported for the time frame before SofS was implemented, since there might be changes in the composition of the population in pilot sites due to the implementation of SofS.

While the baseline populations for the pilot sites and the comparator LAs for EQ3a, EQ3c and EQ3d seem very similar, the populations for the remaining evaluation question differs by individual characteristics. Pilot sites in EQ3b exhibit a much higher rate of re-referrals prior to the settling-in period than comparator LAs.

Table A10.3: Weighted pre-treatment summary statistics for EQ3 a–b broken down by pilots and comparators

Variables	EQ3a Pilot sites	EQ3a Comparator LAs	EQ3b Pilot sites	EQ3b Comparator LAs
Outcome	27	27	27%	16%
Gender – male	49%	49%	49%	49%
Gender – female	47%	47%	49%	49%
Gender – not recorded/unborn	-	-	1%	1%
Gender – missing	-	-	1%	1%
Age	6.34	6.43	8.32	8.31
Disabled	8%	8%	10%	10%
Academic year – secondary school	19%	19%	32%	32%
Academic year – primary school	39%	39%	39%	39%
Academic year – before school age	41%	41%	27%	27%
Academic year – missing	1%	1%	1%	1%
Low income (measured by FSM) – no	12%	12%	16%	16%
Low income (measured by FSM) – yes	35%	35%	29%	29%
Low income (measured by FSM) – below school age so no recorded	39%	39%	25%	25%
Low income – missing	14%	14%	30%	30%
Ethnicity – any other ethnic group	1%	1%	-	-
Ethnicity – Asian	2%	2%	1%	1%
Ethnicity – Black	10%	10%	1%	1%
Ethnicity – missing	11%	11%	5%	5%

Variables	EQ3a Pilot sites	EQ3a Comparator LAs	EQ3b Pilot sites	EQ3b Comparator LAs
Ethnicity – mixed	6%	6%	2%	2%
Ethnicity – unclassified	-	-	9%	9%
Ethnicity – White	69%	69%	81%	81%
Main need – not stated	NA	NA	NA	NA
Main need – abuse or neglect	NA	NA	NA	NA
Main need – child's disability/illness	NA	NA	NA	NA
Main need – parental disability/illness	NA	NA	NA	NA
Main need – family in acute stress	NA	NA	NA	NA
Main need – family dysfunction	NA	NA	NA	NA
Main need – socially unacceptable behaviour	NA	NA	NA	NA
Main need – absent parenting	NA	NA	NA	NA
Main need – cases other than children in need	NA	NA	NA	NA
Main specification (which determines sample size)	Fixed effects	Fixed effects	Linear	Linear
Number of observations in this pre-treatment group	8,387	54,570	16,546	22,930
Number of observations of treatment and comparator LAs over all periods	135,323	135,323	72,807	72,807

Source: Office for National Statistics - National Pupil Database (April 2008-March 2019). Population as described above for each evaluation question. Numbers with '-' are negligible and/or supressed due to statistical disclosure reasons. Percentages are rounded to the nearest whole number and so categories may add to greater than 100%. NA is stated were the variable was not used. Where all values in a row are NA or '-', the row is omitted. All summary statistics are weighted statistics.

Table A10.4: Weighted pre-treatment summary statistics for EQ3 c–d broken down by pilots and comparators

Variables	EQ3c Pilot sites	EQ3c Comparator LAs	EQ3d Pilot sites	EQ3d Comparator LAs
Outcome	1%	1%	39%	34%
Gender – male	48%	48%	52%	52%
Gender – female	52%	52%	48%	48%
Age	10.02	10.06	3.78	3.80
Disabled	4%	4%	25%	25%
Academic year – secondary school	39%	39%	8%	8%
Academic year – primary school	54%	54%	27%	27%
Academic year – before school age	8%	8%	65%	65%
Low income (measured by FSM) – no	41%	41%	2%	2%
Low income (measured by FSM) – yes	42%	42%	11%	11%
Low income (measured by FSM) – below school age so not recorded	3%	3%	63%	63%
Low income – missing	14%	14%	24%	24%
Ethnicity – Asian	-	-	1%	1%
Ethnicity – Black	-	-	4%	4%
Ethnicity – missing	13%	13%	-	-
Ethnicity – mixed	2%	2%	10%	10%
Ethnicity – unclassified	21%	21%	-	-
Ethnicity – White	63%	63%	84%	84%
Main need – abuse or neglect	NA	NA	78%	84%
Main need – child's disability/illness	NA	NA	1%	1%
Main need – parental disability/illness	NA	NA	4%	2%
Main need – family in acute stress	NA	NA	4%	4%

Variables	EQ3c Pilot sites	EQ3c Comparator LAs	EQ3d Pilot sites	EQ3d Comparator LAs
Main need – family dysfunction	NA	NA	11%	8%
Main need – socially unacceptable behaviour	NA	NA	-	1%
Main need – absent parenting	NA	NA	1%	1%
Main regression (which determines sample size)	Logistic	Logistic	Linear	Linear
Number of observations in this pre-treatment group	1,777	854	2,213	4,956
Number of observations of treatment and comparator LAs over all periods	22,529	22,529	11,013	11,013

Source: Office for National Statistics – National Pupil Database (April 2008-March 2019). Population as described above for each evaluation question. Numbers with '-' are negligible and/or supressed due to statistical disclosure reasons. Percentages are rounded to the nearest whole number and so categories may add to greater than 100%. NA is stated were the variable was not used. Where all values in a row are NA or '-', the row is omitted. All summary statistics are weighted statistics.

Analysis results

The tables below detail the results of the primary analysis using different specifications. Levels of statistical significance are indicated by stars. We discuss the results for each evaluation question below the corresponding table.

Table A10:5 details the test results for the Breusch-Godfrey and Hausman tests as well as for the parallel trends test of insignificant leads. Note that for all 4 evaluation questions, the F-tests for joint significance (regressing a binary indicator of whether the individual occurs more than once in the dataset on the outcome of interest and covariates) were highly significant, indicating that individuals that occurred multiple times in the sample differed from those that occurred only once.

Table A10:5: Test statistics for various tests conducted for each evaluation question

Test	Null hypothesis	EQ3a	EQ3b	EQ3c	EQ3d
Breusch-Godfrey test	No serial correlation of order 1 in the errors	5.29*	10.19**	0.96	10.28**
Hausman test	No correlation between the unique errors and the independent variables	55.67***	14.26***	3.85***	2.87***
Number of insignificant leads	Parallel trends test – the majority of treatment leads should be insignificant for parallel trends assumption to hold	0 out of 5 leads are insignificant	4 out of 5 leads are insignificant	2 out of 2 leads are insignificant[5]	4 out of 4 leads are insignificant

Source: Office for National Statistics – National Pupil Database (April 2008–March 2019)

Duration of assessments: Primary analysis

The assumptions of our statistical model were not met for EQ3a, not allowing a causal interpretation of the results. While the identifying assumption was met for the linear model, all 5 of the DiD leads were significant at the 5 per cent level when using a fixed effects model, which calls into question the parallel trends assumption. The data are of sufficient quality (we have no major concerns about identifying the population or the outcome).

Since the Breusch-Godfrey and Hausman tests were statistically significant at the 5 per cent level (see table A10:5 above) and the number of repeated observations is larger than 15 per cent, we chose a fixed effects regression as our model specification. We also examined robustness of the results to alternative model specifications (linear, random effects). While the main results for the primary analysis suggest a significant increase in the duration of assessments through SofS (p=0), the direction of the effect is sensitive to the model specification and the failure of the identifying assumption does not allow a causal estimation of the impact.

[5] Note that this assessment is based on the linear rather than the logistic regression, since the logistic regression specification suffers from convergence issues when introducing individual DiD dummies.

Table A10.6: DiD regression table – estimating the impact of SofS on the duration of assessment (primary analysis)

R^2 =0.06	Fixed effects model	Lagged treatment	Excluding Pilot 2	Excluding Pilot 4	Alternative cut-off
(Intercept)	N/A	N/A	N/A	N/A	N/A
Post settling-in period	9.8***	N/A	6.09***	4.14**	11.92***
Pilot site	N/A	N/A	N/A	N/A	N/A
DiD	2.99***	N/A	0.12	-5.05**	10.82***
DiD in year t	N/A	-0.67	N/A	N/A	N/A
DiD in year t+1	N/A	2.94***	N/A	N/A	N/A
DiD in year t+2	N/A	-4.91***	N/A	N/A	N/A
DiD in year t+3	N/A	-2.83**	N/A	N/A	N/A

Source: Office for National Statistics – National Pupil Database (April 2008 – March 2019). Population: all children assessed. N= 252,169, 5 pilot sites, 5 comparator sites.

Asterisks indicate p-values: *p<0.1, **p<0.05, ***p<0.01

Duration of assessments: Sensitivity analysis

When excluding Pilot 2 who had support from external consultants for assessments from the analysis[6], the effect is no longer significant (p=0.91). Pilot 4 started using SofS prior to the Innovation Programme and the qualitative work by Dr Mary Baginsky suggested that the experience of SofS in Pilot 4 may not be representative of the experience of the other pilots. We excluded it from the analysis to check whether the results were robust to its exclusion. When excluding Pilot 4, the fixed effects DiD coefficient turns negative and significant (p=0.02). Together with the analysis when excluding Pilot 2, these sensitivity checks suggest that the results are primarily driven by these 2 Pilots and cannot be interpreted as average results.

We excluded assessments with a duration above 76 working days because a very long assessment is likely to be an administrative error. We chose this value by the interquartile deviation method, which chooses the threshold based on the distribution of the data: values which are greater than the upper quartile plus 1.5 times the interquartile

[6] Pilot 1 was not included in the main analysis because the closeness of the matches was not sufficient.

range are discarded.[7] However, because it was so closely related to our outcome, we tested the sensitivity of the results to this threshold. We chose a higher threshold of 310 working days based on visual inspection of the distribution of durations to identify visual breaks suggesting outliers. The magnitude of the effect of SofS was considerably higher when the threshold was 310 days with the fixed effect DiD coefficient jumping to 10.82.

Duration of assessments: Secondary analysis

There is a significant decrease in the proportion of assessments conducted within the same day (by 3.94 percentage points, p=0.038), which may help partially explain the increase in the duration of assessment. Parallel to the main analysis, the coefficient changes and becomes positive when excluding Pilot 2 and Pilot 4, respectively (p=0.02 when excluding Pilot 2, p=0 when excluding Pilot 4).

As expected, the impact of SofS on the duration of assessment is of a higher magnitude when SofS is more embedded ($\beta\_3\ (E\_s \cdot T\_t\ )$=18.84, p-value=0) or better delivered ($\beta\_3\ (D\_s \cdot T\_t\ )$=20.17, p-value=0). The magnitude of the effect is larger than when excluding comparator LAs with similar practice models ($〚\ \beta〛\ \_5\ (\ 〚PM〛\ \_(PM = neither, S) \cdot T\_t\ )$=-11.38, p-value=0.003, (note that the coefficient is negative but the direction of the effect hasn't changed, simply the base category). SofS significantly increased the duration of assessments for pilot sites with either low or high Ofsted ratings ($\beta\_4\ (SS\_s \cdot T\_t\ )$=6.64, p=0.001, $\beta\_7\ (SS\_s \cdot T\_t \cdot O\_st\ )$=19.74, p-value=0), and the magnitude and significance of the triple difference coefficient suggests that the effect was of a statistically significant larger magnitude for those with high Ofsted ratings.

[7] The interquartile range (IQR) is a measure of variability, equal to the difference between the 75th and 25th percentiles.

Table A10.7: DiD regression table – estimating the impact of SofS on the duration of assessments (secondary analysis)

Variables	Same day	Quality of delivery	Embeddedness of SofS	Alternative practice models	Ofsted
(Intercept)	N/A	N/A	N/A	N/A	N/A
Post settling-in period	-0.01	7.37***	7.91***	18.85***	2.37
DiD (pilot LA * post settling-in period)	-0.04**	N/A	N/A	N/A	6.64***
Quality of delivery * post-settling-in period	N/A	20.17***	N/A	N/A	N/A
Embeddedness score * post settling-in period	N/A	N/A	18.84***	N/A	N/A
Comparator LA with a similar practice model * post settling-in period	N/A	N/A	N/A	-9.14***	N/A
Comparator LA without a similar practice model * post settling-in period	N/A	N/A	N/A	-11.38***	N/A
Ofsted rating	0.05***	-5.99***	-6.25***	-5.05***	-8.44***
Pilot LA * Ofsted rating	N/A	N/A	N/A	N/A	-9.56**
Post settling-in period * Ofsted rating	N/A	N/A	N/A	N/A	3.95**
Pilot LA * post settling-in period * Ofsted rating	N/A	N/A	N/A	N/A	19.74***

Source: Office for National Statistics – National Pupil Database (April 2008–March 2019)

Population: All children assessed. N= 252,169, 5 pilot sites, 5 comparator sites. Asterisks indicate p-values: *p<0.1, **p<0.05, ***p<0.01

In summary, we see a considerably mixed picture of the impact of SofS on the duration of assessments. The identifying assumption is not met, impeding the estimation of a causal effect. Whilst there doesn't appear to be much of a differential impact of SofS on the duration of assessments (with the exception of Ofsted ratings), the sensitivity to model specification and to excluding particular pilots from the analysis means that we cannot be confident in the direction, magnitude or significance of the effect of SofS on the duration of assessments.

Likelihood of re-referral: Primary analysis

The assumptions of our statistical model were met for EQ3b (4 of the 5 leads were insignificant at the 5 per cent level) allowing a causal interpretation of the results but we have data quality concerns which makes it more likely that we cannot precisely estimate the impact of SofS on the outcome as discussed above. Although the Breusch-Godfrey test statistic is significant, the number of observations that would be used in a fixed effects regression would be less than 15 per cent of the original population and the summary statistics of the entire population and the fixed effects sub-population differ starkly prior to the introduction of SofS (see table A10.3). For this reason, our chosen specification is the linear model. The linear regression suggests that SofS decreased the probability of re-referrals by 9.78 percentage points ($p = 0.0013$). The lagged treatment variables suggest that the effect of SofS is largest in the last year of the observation period (2018/19). We do not deem the evidence to be of moderate or high strength because we were not able to fully account for serial correlation and we have data quality concerns. Pilot sites also have a lower rate of re-referrals prior to the settling-in period than comparator local authorities.

When excluding Pilot 1 and Pilot 2 from the overall sample, the DiD coefficient remains highly significant, of a similar magnitude and of the same direction. Pilot 4 was not included in the main analysis since we could not find comparator LAs with convincing parallel trends.

Table A10.8: DiD regression table – estimating the impact of SofS on the likelihood of a re-referral within 6 months (primary and sensitivity analysis)

R^2 =0.045	Linear model	Fixed effects model	Lagged treatment	Excluding pilots 1 and 2
(Intercept)	0.11**	N/A	0.15***	0.57
Post settling-in period	0.17***	0.47***	N/A	0.06*
Pilot LA	0.04*	N/A	-0.02	0.11*
DiD	-0.1***	0.09	N/A	-0.13***
DiD in year t	N/A	N/A	-0.04	N/A
DiD in year t+1	N/A	N/A	-0.07	N/A
DiD in year t+2	N/A	N/A	-0.13**	N/A

Source: Office for National Statistics - National Pupil Database (April 2008 – March 2019)

Population: all referrals NFA'd at referral or assessment stage. N=72,807, 4 pilot sites, 4 comparator sites. Asterisks indicate p-values: *p<0.1, **p<0.05, ***p<0.01

Likelihood of re-referral: Secondary analysis

The coefficients become insignificant when the pilot dummy is substituted by a scale of the quality of delivery and of the self-reported embeddedness score. Again, this suggests that the scores did not add explanatory value but rather noise, diluting the average effect of SofS on the local authorities. The introduction of a triple interaction term with Ofsted ratings also seems to only further dilute the results and does not seem to add any further insights into subgroup dynamics.

When comparing pilots to just comparator LAs without similar practice models, the coefficient is of a much larger magnitude than the standard DiD coefficient ($〚\beta〛\_5$ ($〚PM〛\_(PM = neither, S) \cdot T\_t$)=0.5, p =0.0004 compared with the coefficient of -0.098 of the main specification). Note that the change in the sign on the coefficient reflects a change in the base category instead of a change in the direction of the effect. This suggests that the main effect potentially underestimates the true treatment effect. However, since the alternative practice model specification does not account for the exact timing when the alternative models were introduced in the respective comparator LAs, the interpretation of β_4 from β_4 ($〚PM〛\_(PM = alternative, S) \cdot T\_t$) as a treatment effect of the alternative model is clouded by the possibility that the dummy actually represents the effect during a time period where the alternative practice model was not implemented during the entire period.

Table A10.9: DiD regression table - estimating the impact of SofS on the likelihood of a re-referral within 6 months (secondary analysis)

Variables	Quality of delivery	Embeddedness of SofS	Alternative practice models	Ofsted
(Intercept)	0.07	0.07	0.15*	0.18
Pilot site	N/A	N/A	N/A	-0.05
Post settling-in period	0.11*	0.12*	-0.11**	0.03
DiD (pilot LA * post settling-in period)	N/A	N/A	N/A	-0.06
Quality of delivery	0.04	N/A	N/A	N/A
Quality of delivery * post-settling-in period	-0.06	N/A	N/A	N/A
Embeddedness score	N/A	0.04*	N/A	N/A
Embeddedness score * post settling-in period	N/A	-0.07	N/A	N/A
Comparator LA with a similar practice model	N/A	N/A	0.06*	N/A
Comparator LA without a similar practice model	N/A	N/A	-0.86***	N/A
Comparator LA with a similar practice model * post settling-in period	N/A	N/A	0.29***	N/A
Comparator LA without a similar practice model * post settling-in period	N/A	N/A	0.5***	N/A
Ofsted rating	N/A	N/A	N/A	-0.14*
Pilot LA * Ofsted rating	N/A	N/A	N/A	0.08
Post settling-in period * Ofsted rating	N/A	N/A	N/A	0.03
Pilot LA * post settling-in period * Ofsted rating	N/A	N/A	N/A	0.1

Source: Office for National Statistics – National Pupil Database (April 2008 – March 2019)

Population: all referrals NFA'd at referral or assessment stage. N=72,807, 4 pilot sites, 4 comparator sites. Asterisks indicate p-values: *p<0.1, **p<0.05, ***p<0.01

Likelihood of a re-referral within 6 months that led to a CPP or LAC plan: Primary analysis

The assumptions of our statistical model were met for EQ4c (2 out of 2 DiD leads were insignificant at the 5 per cent level. The data quality is of insufficient quality which makes it more likely that we do not accurately estimate the impact of SofS on the outcome as discussed above.

We do not find any moderate or high strength evidence that the implementation of SofS affected the likelihood of a re-referral that led to a CPP or the child being looked-after within 6 months of the re-referral date. The main analysis uses a logistic regression specification since the Breusch-Godfrey test yielded insignificant results and the incidence of the outcome is low. Although the results are negative and significant, the small number of local authorities (2 comparator LAs, 2 pilot sites) for which we could find suitable matches in parallel trends and for which we have enough observations after the settling-in period raises concerns with respect to the robustness of the results. In addition, the same data quality concerns as for evaluation question 3 apply. Because of the small number of local authorities in the main analysis, we did not run sensitivity analysis or secondary analysis for EQ3c.

Table A10.10: DiD regression table - estimating the impact of SofS on the likelihood of a re-referral within 6 months that lead to a CPP or LAC (primary analysis)

R^2 =0.01	**Logit model**	**Lagged treatment[8]**
(Intercept)	-18.23***	-0.03
Post settling-in period	-1.43	N/A
Pilot LA	1.78	-0.01
DiD	-6.28**	N/A
DiD in year t	N/A	-0.11*
DiD in year t+1	N/A	-0.12*
DiD in year t+2	N/A	-0.12*

Source: Office for National Statistics – National Pupil Database (April 2008–March 2019)

Population: all referrals NFA'd at referral or assessment stage. N=22,529, 2 pilot sites, 2 comparator sites. Asterisks indicate p-values: *p<0.1, **p<0.05, ***p<0.01

[8] The lagged treatment effects were determined employing a linear model specification, since the maximum likelihood estimation did not converge.

Likelihood of kinship care: Primary analysis

The assumptions of our statistical model were met for EQ3d (all 4 of the DiD leads were insignificant at the 5 per cent level) allowing a causal interpretation of the results. The main analysis suggests that SofS decreased the probability of a child receiving kinship care instead of non-kinship care in pilots as compared with comparator LAs. The coefficient of the linear regression suggests a 12.63 percentage point decrease in the probability of kinship care for LAC during the first 12 months of a period of care (p=0.03). While the Breusch-Godfrey test is significant, the share of observations that would be included in a fixed effects regression would be below 15 per cent of the original sample and we thus do not discuss any multilevel regressions.[9] Pilots 1 and 4 were not included in the main analysis because the closeness of the matches was not sufficient, and hence could not be excluded for the sensitivity analysis. When excluding Pilot 2 from the analysis, the coefficient remains negative albeit of a smaller magnitude, but the result is no longer significant (p=0.29). This suggests that the decrease in kinship care rates might largely be driven by Pilot 2.

Table A10.11: DiD regression table: Estimating the impact of SofS on the likelihood of kinship care – primary and sensitivity analysis

R^2 =0.05	Linear model	Lagged treatment	Excluding pilot 2
(Intercept)	0.35**	0.31**	0.33**
Post settling-in period	0.22***	N/A	0.15***
Pilot LA	0.08**	0.07	0.09**
DiD	-0.13**	N/A	-0.07
DiD in year t	N/A	-0.16***	N/A
DiD in year t+1	N/A	0.01	N/A
DiD in year t+2	N/A	0.01	N/A

Source: Office for National Statistics – National Pupil Database (April 2008 – March 2019) Population: LAC during the first year of a period of care. N=11,013, 6 pilot sites, 7 comparator sites. Asterisks indicate p-values: *p<0.1, **p<0.05, ***p<0.01

[9] We would be able to run random effects regressions using the entire sample, but given the significant Hausman Test results and the risk that the random effects assumptions do not hold, we refrain from comparing random effects with the main regression without a meaningful fixed effects estimator to compare to.

Likelihood of kinship care: Secondary analysis

The analysis of the quality of delivery and of the self-reported embeddedness score does not suggest that the treatment effect differs strongly by the respective scores: the coefficients are of a smaller magnitude and become insignificant.

Accounting for the use of alternative practice models increased the magnitude of the effect from 13 percentage points to 17 percentage points (p=0.029, the coefficient is negative because the base category is different) if comparing local authorities with no alternative practice model and pilot sites using SofS. However, the analysis does not account for the exact introduction dates of the alternative practice models and should thus be treated with caution.

The results from the triple DiD regression suggest no differential effect of SofS by Ofsted rating. While the DiD coefficient is negative and significant (DiD=-0.2, p=0.0002), the triple DiD coefficient is not significant (p=0.14), suggesting that there are no significant differences between local authorities with higher Ofsted ratings and those with lower Ofsted ratings regarding the impact of SofS on the probability of kinship care.

Please note that the analysis only takes into account the first 12 months of a child's period of care, and consequently only considers any placements during the first 12 months of a period of care or special guardianship orders that are appointed within 12 months of a child beginning a period of care.

Table A10.12: DiD regression table - estimating the impact of SofS on the likelihood of kinship care compared with non-kinship care (secondary analysis)

Variables	Quality of delivery	Embeddedness of SofS	Alternative practice models	Ofsted
(Intercept)	0.37***	0.38***	0.45***	0.39***
Pilot LA	N/A	N/A	N/A	0.03
Post settling-in period	0.19***	0.18**	0.04	0.17**
DiD (pilot LA * post settling-in period)	N/A	N/A	N/A	-0.2***
Quality of delivery	0.11*	N/A	N/A	N/A
Quality of delivery * post-settling-in period	-0.11	N/A	N/A	N/A
Embeddedness score	N/A	0.07	N/A	N/A
Embeddedness score * post settling-in period	N/A	-0.07	N/A	N/A

Variables	Quality of delivery	Embeddedness of SofS	Alternative practice models	Ofsted
Comparator LA with a similar practice model	N/A	N/A	-0.13**	N/A
Comparator LA without a similar practice model	N/A	N/A	0.01	N/A
Comparator LA with a similar practice model * post settling-in period	N/A	N/A	0.12***	N/A
Comparator LA without a similar practice model * post settling-in period	N/A	N/A	0.17**	N/A
Ofsted rating	N/A	N/A	N/A	-0.04
Pilot LA * Ofsted rating	N/A	N/A	N/A	0.07
Post settling-in period * Ofsted rating	N/A	N/A	N/A	0.1
Pilot LA * post settling-in period * Ofsted rating	N/A	N/A	N/A	0.16

Source: Office for National Statistics – National Pupil Database (April 2008 – March 2019)

Population: LAC during the first year of a period of care. N=11,013, 6 pilot sites, 7 comparator sites. Asterisks indicate p-values: *p<0.1, **p<0.05, ***p<0.01

Conclusion

The results from the individual-level DiD analysis found no clear evidence for the duration of assessments, and a decrease in kinship care where an increase was expected. The lack of robustness of the results to sensitivity checks and data quality concerns did not allow us to draw firm conclusions about re-referrals and re-referral followed by case escalation. Overall, we would conclude that there is an absence of moderate or high strength evidence of a positive effect of SofS on the outcomes of interest.

The secondary analysis shows that there does not seem to be a differential impact of SofS by the quality of delivery or the self-reported embeddedness of SofS across all pilot sites. The magnitude of the found effects might be underestimating the true effect. Controlling for similar practice models used in comparator LAs, resulted in coefficients of a higher magnitude than in the main specifications. The analysis of alternative

practice models lacks the exact implementation date of these alternative models, so that these results are of an indicative nature only.

This builds on the findings of no evidence in Round 1 of the Innovation Programme, the systematic review from Sheehan et al. (2018) and the narrative review by Baginsky et al. (2019). The outcomes were chosen based on the outcomes which showed most promise during the analysis of Wave 1 and based on MTM's theory of change so we would have expected the most significant, positive impacts of SofS to occur within this set of outcomes.

Appendix 11: Cost study

Methodology

In December 2018 a short survey was sent to all of the pilot authorities (10 at that point) and a reply was received from 8 of them. This survey examined funding, direct and indirect expenditure, and management time spent on SofS.

A follow-up survey was sent in November 2019 to which 6 (of the 9) pilots responded. This survey covered SofS staffing and ongoing costs.

Staffing and training costs

The majority of expenditure was on staff working directly on SofS implementation with authorities reporting that around 75 per cent of the overall spend on SofS was on staffing costs (including those involved in project management). Authorities varied in the number of staff employed to oversee SofS although most reported having some form of 'Project Lead' and 'Practice Development Lead'. Of the 6 authorities that provided detailed staffing information, a total of 21 staff were employed with salaries ranging from £29,000 to £67,000 (median £41,000).

The other large cost area was training, although the variation in spend proportions in this area (between 5% and 31% of total spend) makes clear that there was no consistent approach in how costs in this area were calculated.

Data provided by MTM showed a total of 24 training sessions attended by 866 staff across the 9 pilots. The courses provided were the 5-day training (176 staff), the family finding training (245 staff) or the targeted 1-day training (445). This equated to a total of 2,550 staff days across Round 2. It was not possible to split the Family Finding data by each pilot, but for the 5-day and 1-day training alone, Pilot 5 lost 405 staff days to training over Round 2. Alongside the staffing costs, pilots were required to cover accommodation, travel and subsistence expenses for all attending staff.

There was also a total of 157 Practice Leader sessions provided across the 9 pilots, although attendee numbers varied and were often the same professionals, so it is not possible to get an accurate idea of the total number of staff involved.

Other direct costs

Other direct costs mentioned by pilot sites included venue and room hire, travel and accommodation expenses, IT Development, administration and external consultants. Most authorities were not able to put accurate figures for many of these although the

mean average estimate (from 4 authorities) for venue hire was nearly £15,000 in 2018/19 and for IT development (from 3 authorities) was over £32,000.

Indirect costs

Identifying indirect or hidden costs accurately is problematic as, by definition, they are costs borne by the organisation which are not listed individually and often come from different budget allocations. When asked about the hidden costs, the most commonly reported items were related to providing backfill for posts when staff went on training, management time (which is dealt with separately below), promotion and communications, and travel and accommodation expenses with 1 site reporting that:

> …we have had a lot of practice leads sessions and as the workers claim mileage via their individual team, I don't think we could easily find out about this hidden cost, or for their time.

Management input for SofS implementation

Authorities were asked about additional management staff input to support SofS implementation, over and above that funded through the Round 2 grant. Management staff time is rarely recorded in this way and sites found it challenging to even provide an estimate of FTE (full time equivalent staff) spent.

Over the 6 sites that provided data on this, the majority of management time was at Service Lead or Principal Social Worker level. Estimates varied wildly between 5 per cent of time in 1 authority to 80 per cent in another, suggesting clear methodological differences in how they were estimating time spent.

At senior management level, 5 of the 6 sites apportioned time to the Assistant Director of Child Services (although again this varied between 65% and 3%) with Directors of Child Services estimated to have spent very little time directly on SofS implementation.

According to the estimates provided by the 6 pilot sites, over the 2-year period, there had been an average of 230 days input from management staff across all grades, within a range of 52–1118 days per site. In Round 1, management time was fairly evenly split between 'senior' and 'middle' managers, whereas in Round 2 the amount of senior time spend has reduced.

Ongoing costs

Authorities were asked to provide information on what factors they considered important to both the successful implementation of SofS and its sustainability over the longer

term. All 6 responding authorities thought staff dedicated solely to SofS were important to implementation and sustainability.

In contrast, 4 of the 6 thought external consultants were important to successful implementation, but only 2 thought this was true for sustainability in the longer term, with 3 saying they were of low importance.

Of the 21 staff reported to be employed on SofS in 6 of the sites, 17 of them were expected to continue to be employed beyond the end of Round 2, with only 1 of those 17 having a temporary contract.

Table A11.1: Costs related to Signs of Safety during Round 2 and expected to continue in future

Type of expenditure	Used during Round 2	Ongoing in future
Training	83%	67%
Venue/room hire	100%	67%
Administration	67%	67%
Updating guidance/forms	67%	50%
Marketing, promotion and communications	83%	83%
IT development	83%	83%
Travel and accommodation expenses	100%	50%

Many of the costs reported for implementation are expected to continue in the future, as shown in Table A11.1, with ongoing training required due to staff turnover and IT development being 2 of the major cost items. Without external funding it is unclear to what extent the observed levels of expenditure are realistic and sustainable over the longer term.

© Department for Education

Reference:

ISBN:

The views expressed in this report are the authors' and do not necessarily reflect those of the Department for Education.

Any enquiries regarding this publication should be sent to us at:
CSC.Research@education.gov.uk or www.education.gov.uk/contactus