Journal of Family Violence (2023) 38:1177–1189
https://doi.org/10.1007/s10896-023-00549-z

ORIGINAL ARTICLE

The Devil's in the Detail: Implementation Fidelity in Evaluating a School-Based Prevention Programme for Children Under 12

Zain Kurdi[1] · Annemarie Millar[2] · Christine Anne Barter[3] · Nicky Stanley[3]

Accepted: 31 March 2023 / Published online: 10 April 2023

Abstract

Purpose This paper examines implementation fidelity (IF) and the underpinning conceptual framework drawing on the evaluation of a UK-wide, manualized child abuse and neglect prevention program for elementary schools. We describe and assess our approach to assessing IF and consider how IF can inform program development.

Method We drew on the literature on program fidelity and critical components of the program evaluated to identify three dimensions of IF: *Coverage, Quality* and *Context*. Data was collected through external observations using systemized observation schedules which were extracted to be scored using scoring protocols for each intervention type. Scores were calculated by two researchers with a random sample cross-checked by a third member of the research team.

Results Observation analysis demonstrated consistency in the coverage of content when delivering assemblies for both younger and older children with at least 76% coverage of content across the assemblies. However, observation analysis revealed greater levels of variability in the delivery of workshops. Material on sexual abuse was less fully covered and children reported that some facilitators lacked confidence and clarity in delivering this material (Stanley et al., 2023).

Conclusion Our results indicate the usefulness of systemized observations in capturing coverage of content, these findings underscore the importance of developing scoring protocols and training observers prior to evaluating program delivery. We highlight the significance of integrating implementation fidelity training for program facilitators and implementers to both assist with monitoring and to maintain quality, despite variations in the actual delivery and setting of the program.

Keywords Implementation Fidelity · Prevention programs · School-based interventions · Measuring effectiveness · Child abuse and neglect

Introduction

This paper examines implementation fidelity (IF) and its underpinning conceptual framework. We consider the role of IF in the delivery and evaluation of school-based prevention programs for children, describing the process adopted to assess and measure IF for the independent evaluation of one widely delivered UK program. The "Speak Out. Stay Safe" (SOSS) child abuse and neglect prevention program was developed by the National Society for Prevention of Cruelty to Children (NSPCC) and was delivered to primary schools across the United Kingdom (UK) by trained staff and volunteers. We report on the process utilized to measure implementation fidelity in a representative sample of participating schools and share our results and reflections on the methodology, offering recommendations for implementing and evaluating similar evidence-based interventions.

✉ Zain Kurdi
z.kurdi@ed.ac.uk

Annemarie Millar
contact@annemariemillar.com

Christine Anne Barter
CABarter@uclan.ac.uk

Nicky Stanley
NStanley@uclan.ac.uk

[1] School of Social and Political Science, University of Edinburgh, 15a George Square, Edinburgh EH8 9LD, UK

[2] Social Work, Queens University Belfast, 6 College Park Ave, Belfast BT7 1PS, UK

[3] Connect Centre, University of Central Lancashire, Preston, Lancashire PR1 2HE, UK

By providing a detailed account of the process, we aim to extend understanding and measurement of implementation fidelity for child abuse and harm prevention programs aimed at young children. We consider three key questions regarding the relationship between IF and program outcomes: what is implementation fidelity; who needs to understand it and who measures it?

Literature Review: What is Implementation Fidelity?

Implementation fidelity (IF) is "The degree to which…programs are implemented…as intended by the program developers" (Carroll et al., 2007, p. 1). Implementation fidelity acts as a potential moderator of the relationship between interventions and their intended outcomes. IF remains a concept that is often ignored and/or poorly understood; for instance there are varying interpretations of what constitutes the core components of fidelity (Carroll et al., 2007; Century et al., 2010; Fixsen et al., 2009; Lynas & Hawkins, 2017a).

There has been a steady increase in the study of IF with several reviews across various disciplines exploring the importance of well-implemented programs in achieving maximum results (Bruhn, Hirsch, & Lloyd, 2015; Durlak and DuPre 2008; Griffith, Duppong Hurley, & Hagaman, 2009; Sanetti, Dobey, & Gritter, 2012). Yet relatively little is understood about the contribution of IF to the effectiveness and retention of key messages from prevention programs on abuse and harm aimed at young children (under 12) despite over three decades of such programs being delivered in schools in both the UK and US (Gubbels et al., 2021; Holloway & Pulido, 2018). Intervention programs are often characterized as a 'black box'; meaning little is known about their functioning (Haynes et al., 2015, p. 2).

Two reviews assessing the role of implementation fidelity within health (Durlak & DuPre 2008) and mental health (Rojas-Andrade & Bahamondes, 2019) programs targeting youth in both community and school settings, found that interventions that are well implemented have effect sizes up to three times larger when compared to poorly implemented interventions. They also state that under "ideal conditions" interventions with high IF can be up to 12 times more effective than poorly implemented ones" (Rojas-Andrade & Bahamondes 2019, p. 341). Identifying minimum desirable thresholds to achieving effectiveness in prevention interventions is essential in order to be able to replicate them in everyday school settings, which require flexibility due to their diverse nature (Sarno et al., 2014).

In the rise of evidence-based interventions, IF is increasingly recognized as an important enabler, ensuring intended intervention outcomes are achieved and unintended consequences are minimized (Fixsen et al., 2021). Recognition of the value and necessity of considering IF will hopefully ensure interventions are not allowed to just happen but are made to happen. Despite a burgeoning in the literature dealing with the measurement of implementation fidelity both across disciplines and within the field of Implementation Science (IS) itself, there remains a lack of consensus on what constitutes core components, and the implications of inconsistent application of methods to ensure fidelity (Gearing et al., 2011; Lynas & Hawkins, 2017a). Identifying critical components within school-based prevention program can ensure minimum thresholds are established to achieve program fidelity despite variations in context, while also allowing for a degree of flexibility during implementation (Bertram, Blase, & Fixsen, 2015; Fixsen et al., 2009, 2021).

There are gaps in knowledge on how specific program components and delivery techniques relate to the effectiveness of child abuse prevention programs (Gubbels et al., 2021; Lynas & Hawkins, 2017b). In order to achieve effective implementation in innovative programs, it is important to look at the drivers of implementation within a "complex human service environment" (Fixsen et al., 2015; Fixsen & Blase, 2020). Human services relate to a wide range of services in which one person interacts with another, such as a teacher with a student, in a way that is intended to be helpful (Fixsen & Blase, 2020). The 'human service' aspect of IF relates to two important components: firstly, the quality of delivery, whereby the facilitator delivering an intervention has been trained with the required skill set to deliver sensitive, new concepts to young children, what is termed the 'enactment of skills' (Lynas & Hawkins, 2017a). Secondly, levels of engagement by the intended recipients of an intervention and the engagement of the organization hosting the intervention, such as a school. A systematic review on the significance of IF on school-based mental health program outcomes concluded that the strongest association between IF and outcomes is students' exposure and receptiveness to the intervention (Rojas-Andrade & Bahamondes, 2019).

Our approach to measuring fidelity was informed by that of other researchers who have developed their approaches in a wide variety of studies (Harn et al., 2013; Bauer et al., 2015; Carroll et al., 2007; Century et al., 2010; Fixsen et al., 2021; Haynes et al., 2015). These authors highlight the importance of identifying the various components or building blocks that make up an intervention, these are also referred to as dimensions (Carroll et al., 2007; Century et al., 2010; Haynes et al., 2015). Implementation efforts are complex due to the multiple interacting levels (between program beneficiaries, stakeholders and implementers), coupled with diversity in the implementation setting (Bauer et al., 2015). The importance of the implementation setting and venue has been highlighted as an integral element

within implementation science: in its definition it includes the 'environmental characteristics in which implementation occurs' (Damschroder et al., 2009). However, most implementation theories in the literature use the term, 'context' or 'setting' more broadly to include indicators for measuring contextual aspects, relating to what Damschroder et al. (2009) describe as inner and outer settings. Although the line between inner and outer settings is not always clear, within the context of a school-based intervention, aspects such as engagement with beneficiaries, characteristics of the facilitators and those involved in the organizational aspects of the program, as well as the underlying theories behind the intervention would be seen as belonging to the outer setting (Damschroder et al., 2009; Fixsen et al., 2021). The inner setting includes the physical space within which the intervention takes place (venue) and the structural and cultural context of the school where it is being delivered.

We identified five dimensions to fidelity; these were: *adherence to the intervention model, exposure or dose, quality of delivery, participant responsiveness;* and *program differentiation* (Carroll et al., 2007; Century et al., 2010; Haynes et al., 2015). In this paper, we share how we operationalized these dimensions while evaluating IF within a school-based child abuse and neglect prevention program (SOSS), described below, as part of a wider evaluation of the program across the UK (Stanley et al., 2023). We hope that sharing and reflecting on our experiences of measuring IF may inform the design of new interventions and future evaluations.

The Speak Out Stay Safe (SOSS) Program

NSPCC's Speak Out, Stay Safe (SOSS) is a child abuse and neglect prevention program delivered in mainstream primary schools[1] across the UK aimed at increasing children's understanding and awareness of abuse and harm and enabling them to identify it and seek help from a trusted adult. During the period when the program was evaluated, the SOSS program was delivered by trained NSPCC staff and volunteers via a 20-minute school assembly for younger children aged 5-7yrs, and a 30-minute school assembly for older children aged 7-11yrs, followed by an interactive one-hour workshop for older pupils only (aged 7–11). All three elements of the program are meant to be delivered by a pair of facilitators. The SOSS is a manualized program with a separate manual and set of presentation slides for each element of the intervention (younger and older children assemblies and workshops) (NSPCC n.d.), the program mascot is a friendly, green speech bubble, called Buddy which emphasizes the need to confide and speak out to a trusted adult. All three elements of the intervention include a presentation with embedded interactive video clips and short animations. The assemblies are short and focused on introducing different forms of child abuse and neglect, the importance of speaking out and confiding in a trusted adult, in addition to raising children's awareness of the Childline (the NSPCC's national helpline for children) number and ways in which it can be accessed. The workshop element is more interactive and longer, lasting an hour; its aim is to explore sexual abuse and neglect, and it relies on facilitator interaction and engagement with the children and school staff present. The expected venue to be used for the delivery of the various elements of SOSS are an assembly or gym hall for the assemblies and a classroom for the workshops.

Each element of the SOSS program is meant to be delivered once within an academic year. The NSPCC aims to deliver the program in schools as often as resources allow, with some schools receiving the program on an annual basis. Since program uptake depends on schools opting in, some schools engage with the program in a more sporadic way. Ideally, children receive the intervention (an assembly) at least once during their time within the earlier stages of primary school and at least once in the form of an extended assembly and workshop during their later stages of primary school (Stanley et al., 2021).

Methods

TESSE Evaluation of the SOSS Program

This paper reports on one element of the TESSE (**T**he **E**valuation of **S**peak Out, **S**tay Saf**E**) (Stanley et al., 2023), evaluation of the SOSS program, focusing on IF. The main evaluation was preceded by a pilot evaluation to assess the tools developed for the main evaluation (Barter et al., 2022). The mixed methods evaluation aimed to examine the program's impact on children's understandings of abuse and harm and their reported help-seeking and to investigate the experiences of program participants (for a fuller account of the evaluation methods see (Stanley et al., 2021). Here we focus on program fidelity which was assessed in order to discover whether the program was delivered as intended.

The integrated process evaluation was conducted in 13 of the intervention schools and included observation of program implementation and fidelity and interviews with 16 teachers and 15 program facilitators. Focus groups were held with a total of 61 children to capture their experiences of both program content and delivery.

[1] In the UK, mainstream schools are the majority of schools; they are not special schools which are schools intended for children with a high level of special education needs or disabilities.

Ethics

The evaluation went through three separate ethical committees and received approval from the NSPCC Ethics Committee, ethics committees at the University of Central Lancashire and the University of Edinburgh. All NSPCC staff and volunteers were asked to consent to observations of assemblies and workshops, all facilitators presenting in process schools that were approached, provided consent for observation (Stanley et al., 2021). All schools, children, teachers and facilitators have been anonymized.

Evaluating the Implementation Fidelity of the SOSS Program

A structured process evaluation was used to assess the IF of the manualized SOSS program in a representative sample of 13 participating schools. The key research question in respect of fidelity was: how does program delivery vary and is this significant? Four additional sub-questions addressed the following: how much variation is there in delivery? Are some elements of the program prioritized or delivered more fully than others? Are there any differences between delivery among the younger and older children? Are there differences in the consistency of assembly and workshop delivery?

Three observation schedules were developed by the research team using the program handbooks for the two assemblies (older and younger children received different assemblies) and the workshop. The observation schedules include the program's essential elements and covered the full range of delivery methods for the two assemblies and the workshop for older children. The observation schedules (see online supplementary material A for an example) captured adherence to content including every aspect of the intervention, with space for qualitative feedback by observers. The observation schedules also incorporated demographic information about the facilitator/s observed, their enactment of skills, such as tone and pace of delivery, familiarity with material and comfort in presenting the material. The suitability of the venue was covered, noting access to suitable technology to present the intervention. Finally, the level of adult engagement from key staff (including both school staff and other adults such as parents present) at the host school and child engagement were captured. A checklist approach was used to record some elements such as level of participation by the children or proportional coverage of a certain element of the program. Throughout the observation schedules, qualitative comments were recorded in an open box that allowed researcher to add their observations of the aspect being assessed.

A three-dimensional IF framework and scoring model was developed based on the five dimensions identified in the literature and the critical components of the SOSS program. We did not include dosage as the intervention was delivered once during the evaluation lifecycle. We condensed the four remaining dimensions further into three, including participant responsiveness and program differentiation as sub-dimensions. These three dimensions: coverage, quality and context are shown in Fig. 1. The coverage dimension was informed by a meeting with program developers and delivery staff who identified the 'essential components' and

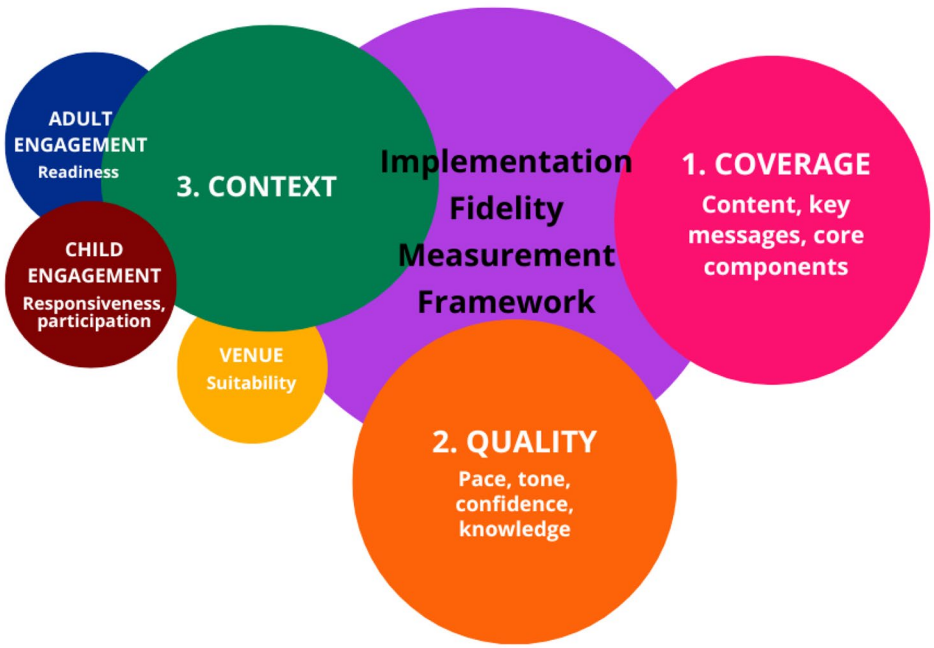


Fig. 1 Key Dimensions contributing to Implementation Fidelity

Table 1 Mean and median observation Scores for Assemblies for the younger children aged 5-7yrs: all scores are out of a potential total of 33.3 and the total score is out of 100

Dimensions	MEAN	MEDIAN
Coverage	**28.5**	**28.1**
Quality	**26.3**	**27.0**
Context	**21.3**	**22.1**
Total	**76.1**	**74.1**

*All three dimensions contribute equally to a percentage score representing our measure of the fidelity of implementation of the Speak Out. Stay Safe program

Table 2 Mean and median observation scores for the older children aged 7-11yrs Assemblies: all scores are out of a potential total of 33.3 and the total score is out of 100

Dimension	MEAN	MEDIAN
Context	**23.0**	**23.6**
Coverage	**30.8**	**31.2**
Quality	**27.2**	**27.7**
Total	**81.2**	**79.0**

*All three dimensions contribute equally to a percentage score representing our measure of the fidelity of implementation of the Speak Out. Stay Safe program

key messages of the program. This allowed us to attribute higher scores to those elements identified as contributing directly towards program outcomes. The quality dimension, also informed by discussion with the program developers, related to facilitators' comfort in relaying and discussing sensitive concepts, levels of confidence and ability to engage children's interest and attention. This dimension covers what Damschrdoer et al. refer to as outer setting; in the context of this study, the outer setting constituted the external facilitators visiting the school (Damschroder et al., 2009). This dimension was considered as important as coverage, as 'enactment of skills' has been highlighted as essential to achieving IF in school-based child sexual abuse prevention programs (Lynas & Hawkins, 2017a). The context dimension included all aspects of inner setting referring to the school space, in addition to stakeholders' (school staff) and beneficiaries' (children) engagement. One element of Damschroder et al's (2009) outer setting was included within the context dimension: that was working technology which, in this case, was as dependent on the school's facilities (projectors, internet connection, working speakers and microphones); as it was on facilitators checking beforehand whether a school was equipped to accommodate delivery requirements.

As this study was conducted across four countries, six researchers undertook the observations. Consistency in recording observations was developed in the piloting stage with training provided for researchers who joined the team at a later stage. The researchers were already familiar to children participating in the evaluation since they had administered baseline surveys in the classroom prior to program delivery. They were introduced to children again during the program's opening session and they then sat down to watch the assembly or workshop alongside the children, so experiencing the same comfort or, in some cases, discomfort as the children experienced. Most observations were conducted by one researcher and lasted for the entire session. In some instances, researchers observed more than one element of the intervention on the same day. On two occasions, two observers observed the same intervention element. Both observation sheets were consolidated by a third member of the research team and the observation was extracted as one.

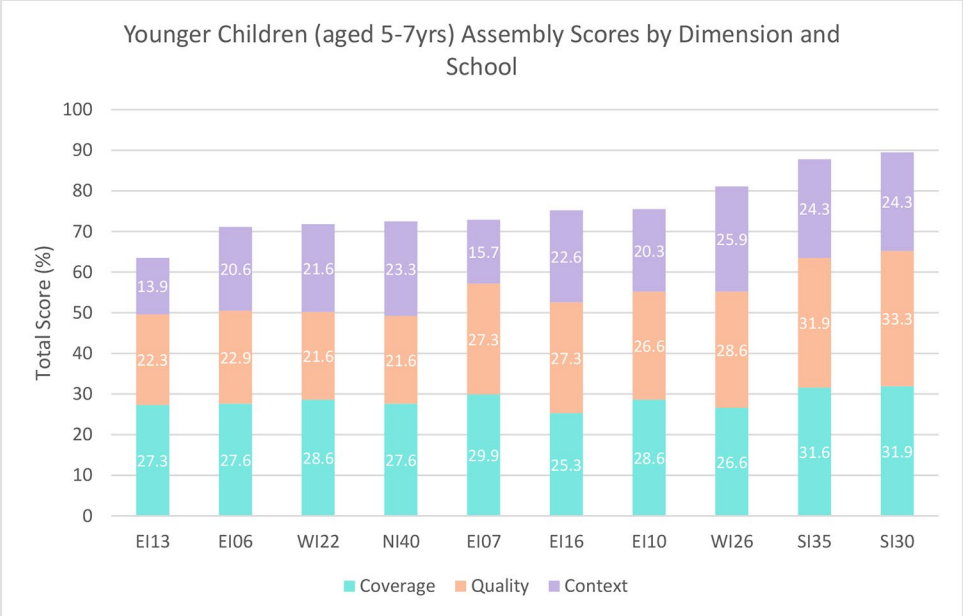
During observations, the researcher situated themselves in the assembly hall or classroom in a position where they could observe the facilitators, school staff and children while remaining apart. This distanced the researcher from the implementation setting and ensured that they were not perceived as part of the program or any of its activities. During workshops, researchers would at times walk between the groups of children who were seated in discussion groups to pick up on their levels of engagement and those of facilitators and school staff.

Analysis

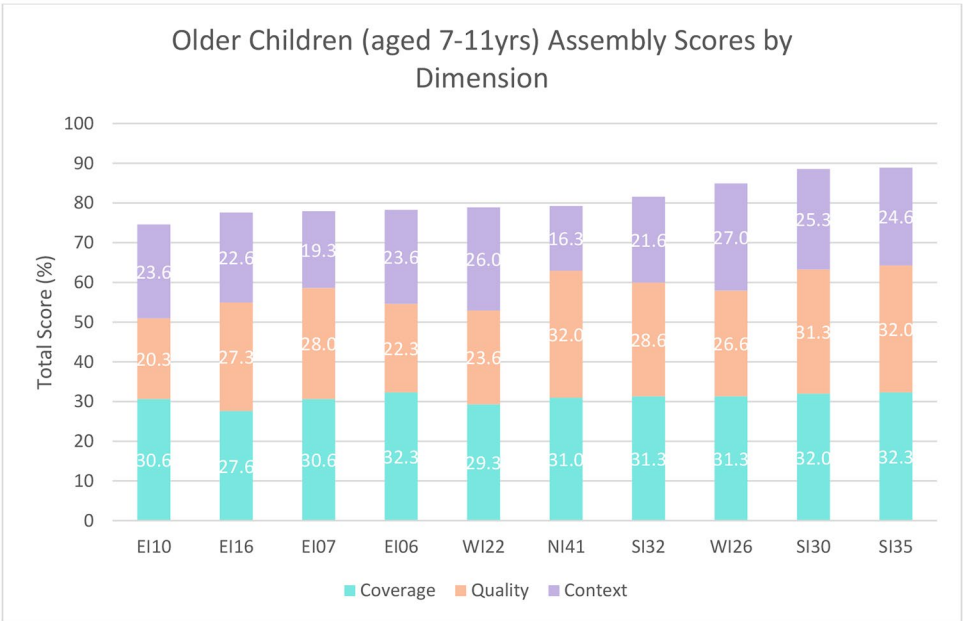
A scoring protocol was used to translate all checklist and binary items into weighted scores, qualitative comments contributed less formally to the score. The concept of *congruence* was used to quantify observed categorical data and analyze qualitative comments. If the comment following a checklist or appraisal reinforced the chosen item and was deemed congruent, an additional score was added; if however, the comment contradicted the chosen item, a proportion of the element's score was deducted. Qualitative comments also contributed to assessment as items of observation in and of themselves. Depending on the area being commented on and its importance to the IF of SOSS, a maximum score was allocated to the comment, with scores given according to the extent to which a comment supported the specific area of implementation. An example of how this was operationalized can be seen in the older children observation and scoring workshop protocol which is included under Supplementary material B (available online).

In the analysis of scores, all three dimensions - *Coverage, Quality* and *Context* - contributed equally to a total percentage score representing IF (with a score closer to 100 per cent indicating that delivery was closer to program guidance and intention). A scoring system was developed for each element of the program. Observation schedules for each of the three intervention types were extracted into a master extraction file and then scored by two researchers.

Graph 1



Graph 2



A sample of these extractions were then checked by a third researcher to ensure inter-rater reliability.

Results

In total, 56 facilitators were observed. Ten observations were completed of each element of the program: 10 assemblies for younger children aged 5–7 years, 10 assemblies for older children aged 7–11 years and 10 workshops.

Comparing Fidelity Scores Across the Three Dimensions

Fidelity scores from the assemblies for both younger and older children showed similar trends and were comparable across all three dimensions. Scores from the older children's (aged 7-11yrs) assemblies were higher across all three dimensions, only marginally higher for the Quality and Context dimensions (see Graph 1 & Graph 2). The highest scores were found under the Coverage dimension for both younger and older children's assemblies.

The range of variability of inter-dimensional scores (difference between the lowest and highest scores across 10 observations) was similar for all three dimensions for both younger children's and older children's assemblies. The Coverage Dimension had the lowest inter-dimensional score in both sets of assemblies, with a score range falling within 6.6 points for younger children and 4.7 points for older children's assemblies. The score range for the Quality Dimension was exactly the same for both set of assemblies, with a score range falling within 11.7 points. The Context Dimension had a slightly higher score range for younger children's assemblies (12 points) observations compared with older children's assemblies (10.7 points).

In the workshops, unlike the assembly scores, the Coverage Dimension scores were lower than the Quality and Context Dimensions. The inter-dimensional score ranges for the workshops (Table 3) were much higher showing much more variability in delivery of workshops. Coverage variability was lowest with a score range falling within 10.6 points. The Context Dimension variability was higher again with a score range falling within 15.5 points. Finally, the variability of scores within the Quality Dimension was the highest (18.3 points) seen both within workshop observation scores and all scores across the assemblies.

These results show consistency in Coverage when delivering SOSS assemblies for both younger and older children with at least 76% of Coverage achieved across the assemblies. The higher Coverage scores for older children's assemblies are explored further below.

Median Coverage and Quality scores differed between faith and non-faith schools. Faith schools had lower Coverage at 59% (65.5% non-faith) and lower quality in delivery at 49% (67% non-faith). However, the Context Dimension scored higher in faith schools (70.1%) compared to non-faith schools (65%).

Coverage Dimension

The coverage of the younger children's assemblies scored a median of 28.1 (Table 1) and older children's (aged 7-11yrs) assemblies scored a median of 31.2 (Table 2). The range of scores under the Coverage dimension for both the younger and older children's assemblies found high levels of consistency in coverage of material with the lowest observed Coverage Score 25.3 (76%) in an assembly for younger children (aged 5-7yrs) delivered in England (Graph 1).

Table 3 Mean and median observation scores for workshops for older children aged 7-11yrs: all scores are out of a potential total of 33.3 and the total score is out of 100

Dimension	MEAN	MEDIAN
Coverage	**19.7**	**20.7**
Quality	**22.8**	**21.4**
Context	**21.6**	**22.5**
Total	**64.2**	**62.3**

*All three dimensions contribute equally to a percentage score representing our measure of the fidelity of implementation of the Speak Out. Stay Safe program

Differences in the delivery of certain elements of the program were observed and might explain the higher coverage scores for older children's assemblies. The Childline Key Information element in the younger children's (aged 5-7yrs) assemblies had the lowest consistency in coverage with an average that was half that of other Coverage in that section. The 'Grownups that may be good to talk to' element had lower average coverage than the other elements in the younger children's (aged 5-7yrs) assembly, this might be due to the interactive nature of this element where children participated and contributed to steering the discussion. If children don't identify certain trusted adults, the facilitator might omit discussion of them.

Digitized program components were consistently delivered with almost half of younger children assemblies scoring the maximum score on the coverage of the video element of the assembly (Sam's Story). In the older children's assemblies, all observations of the video element of the program scored 100% on coverage. The lowest average coverage among the older children's (aged 7-11yrs) assembly observations was the 'Sources of Help/Emptying the Sack": this could again be due to the participatory/interactive nature of this element.

The older children's assembly includes more digitized content and is 10 min longer, allowing more time for the coverage of content and participation: this could explain why the Coverage Dimension scores for the older children were higher than those for the younger children.

Workshop observations highlighted that entire components of program content were sometimes omitted. An example of this was the assembly recap section at the start of a workshop. The purpose of the assembly recap is to remind the children about the key messages received during the assembly. On some occasions, some children will receive the assembly and workshop on the same day, alternately, facilitators may return within a week or so to deliver the workshop, making the assembly recap pertinent. In two of the ten observations, the Assembly Recap was left out by the facilitator. These two schools were not the only schools where delivery of both the assemblies and workshops took place on the same day and does not explain why this element was excluded.

On average, only 51% of the specified coverage of the section on abuse topic A (sexual abuse) (which included 5 different elements) was covered in the workshops. Since the workshops rely on children's engagement and participation, having the necessary confidence, tools and skills to handle small group dynamics is integral to delivery of these core

program elements. Some children participating in focus groups noted that facilitators seemed "scared" when the topic of sexual abuse was introduced.

The lowest coverage score in this section of the workshop was for the school where a volunteer delivered the workshop alone and was just 17.6%. The workshop was not delivered in an appropriate venue, this workshop scored lowest on the sub-dimensional venue score and children as young as five kept coming in and out of the hall where the workshop was being delivered, leading to start-stop style of delivery. Behavior management was also another issue for delivery.

The highest coverage score of 71.7% for abuse topic A was given to a workshop delivered jointly between NSPCC staff member and a volunteer. The segment was delivered by the NSPCC staff in its entirety, the facilitator was thorough and reinforced key messages for children using examples:

> *"Facilitator gives example of snogging [extended kissing] on TV and uses children's discomfort as an example of why it is not an appropriate type of kiss for children to receive" (Workshop Observation, 3).*

Quality Dimension

Of the 56 facilitators observed, 19 were observed delivering assemblies for younger children, 18 delivering assemblies for older children, and 19 delivering workshops. Most facilitators delivered more than one element of the program. The age range of observed facilitators started at 25 years of age and went up to 74 years of age with the majority of facilitators falling in the 45–64 age range. Most (80%) of facilitators were older women and trained volunteers, this is in keeping with the broad picture of SOSS program facilitator demographics. Five NSPCC staff were observed delivering the intervention alongside volunteers. All workshops were delivered by a pair of facilitators except on one occasion where a female volunteer delivered the workshop alone.

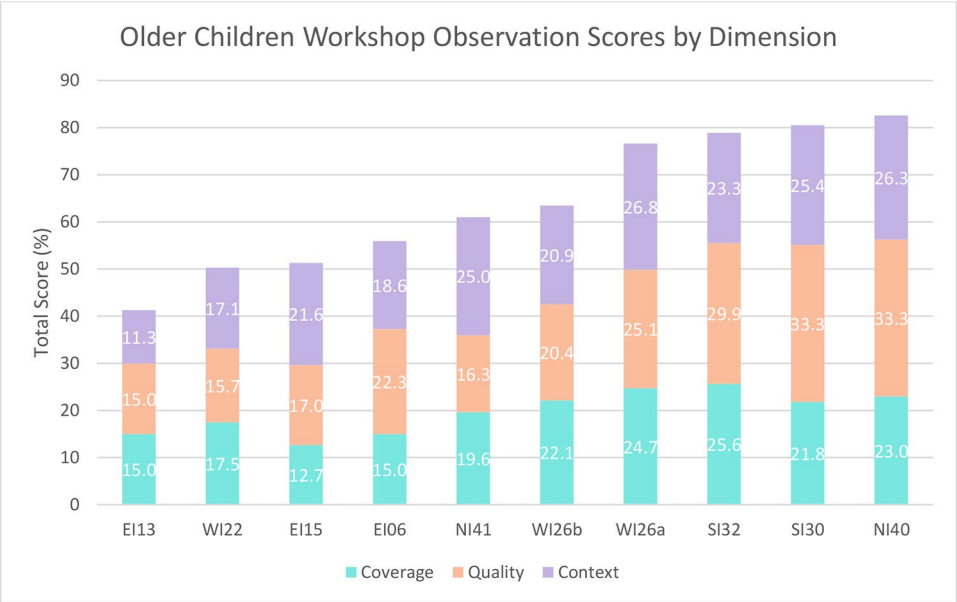
There was more inconsistency in the delivery and implementation of workshops (see Graph 3) than in assemblies. The participatory and interactive nature of workshops presented a more complex environment for facilitators to navigate, despite the student group being smaller. Workshop facilitators, all but one of whom were volunteers, seemed to struggle at times to manage discussions or tricky questions. In the less structured workshops, facilitators' levels of experience, training and confidence come to the fore: this may be an area for improvement.

Context Dimension

Venue

An important aspect of the context dimension is the delivery venue which is part of the inner setting (Damschroder et al., 2009). Here, inner setting refers to the school space. The venue sub-dimensional score was one of three scores contributing to the Context Dimension, a venue was deemed unsuitable if it was noisy, busy and didn't provide the required setting for the element being delivered. Twenty per cent of observed assemblies for younger children, 50% of observed assemblies for older children and 60% of the workshops received a full score on venue suitability and appropriateness. The venue scores for the younger children

Graph 3



receiving the assembly were lower than average: children were observed sitting on hard, uncomfortable floors or in hot, crowded or noisy rooms. This suggests that rethinking the contextual (environmental) needs of younger children receiving the program would be timely, and importantly would be consistent with a child rights-informed approach (Lundy, 2019).

Child Engagement

The workshops are included in the program with the aim of offering older children opportunities for more active and in-depth learning. The results clearly demonstrate the importance of the human dimension for implementation fidelity: these are not considered critical components but are nonetheless; essential to program outcomes (Century et al., 2010). There was considerable variability regarding the extent to which the Buddy kit element of the workshop was covered and the time allocated to it varied greatly. This impacted on children's opportunities for participation:

> *"Children appear to be excited to receive their Buddy kit. Facilitator 1 emphasizes that no one is going to look at their workbook or mark it. However, after approximately one minute, children are told to put their workbooks away." (Workshop* Observation, 106).

Adult Engagement

The adult engagement score focused on presence of school staff and their contribution to program delivery. We recorded whether school staff remained in the assembly or workshop after accompanying the children in and the level of their engagement throughout and at the end of these sessions when some school staff made closing remarks.

There was no statistically significant correlation between adult engagement and child engagement in assemblies for either younger or older children. A Pearson's Correlation test was run to establish whether adult engagement scores were predictive of child engagement scores; the results were not statistically significant ($r=-0.56$, $p=0.096$) and neither a positive or negative correlation was evident.

Discussion

Our aim was to systematically measure IF in a primary school-based preventive intervention; the process has allowed us to develop a detailed account of our observation, scoring and analysis techniques. Measuring fidelity extends beyond adherence to program content; the delivery setting and staff engagement play a key role in creating a suitable space for the program's key messages to be received and absorbed. IF is therefore a multidimensional construct and other studies emphasize the importance of both facilitator characteristics (e.g. experience of delivering program, training, confidence delivering different aspects of program), and participant behavior (Berkel et al., 2011; Durlak & DuPre, 2008).

School readiness and engagement are also concepts that have been highlighted in previous literature on integrity in evidence-based interventions: "the majority of preventive interventions are conducted in schools; their success will depend on the recognition by school administrators, teachers, and other personnel of their utility and practicability within an already full school schedule" (Dane & Schneider, 1998, p. 24; Hansen et al. 1991). Delivery of preventive interventions is dependent on the setting in which they are delivered and implementers will often have to contend with myriad obstacles to the fidelity of program delivery. A review on the influence of implementation on program outcomes, noted over 20 contextual factors that influenced implementation and outcomes (Durlak & DuPre, 2008). Moreover, as this study found, the quality of delivery also impacts on delivery of key messages and this may contribute to program effectiveness or lack of it. Measuring IF in any intervention is essential to realizing outcomes of evidence-based interventions (Cutbush et al., 2017).

Implementation fidelity therefore consists of multiple contributing and overlapping dimensions which need to be assessed and measured. Figure 1 shows the three main dimensions we used to encapsulate the critical components of the SOSS program and the overlapping sub-dimensions measured under Context. However, this figure suggests that all dimensions contribute equally to implementation fidelity and this may not necessarily be the case.

To ensure IF is measured within evidence-based programs and interventions, developers need to further develop their intervention manuals by establishing and highlighting the essential components of an intervention. They also need to recognize the inevitable contextualization of any program and the need for a degree of flexibility and responsiveness to local conditions to be built into the design (Haynes et al., 2015). This is certainly true for a program delivered as widely as SOSS: in the 2018/19 academic year, the program was delivered to approximately 1.8 million children in 8000 schools in the UK and Channel Islands (personal communication from the NSPCC). Therefore, some variation in delivery is inevitable and has to be built into the implementation strategy. Our study used qualitative data to capture such variations in delivery and response to local conditions, highlighting the question about who needs to understand the value and uses of fidelity.

For school-based programs, it is vital to acknowledge that implementation fidelity of preventive programs cannot solely focus on coverage (program content) but as the findings here and in previous work (Bertram et al., 2015; Cutbush et al., 2017; Fixsen et al., 2021; Haynes et al., 2015) show, attention must also be given to the role of 'quality of delivery' - the enactment of skills by facilitators - and the 'context'. It is possible that these dimensions of IF are related to one another but will most likely operate in different ways to influence outcomes (Durlak & DuPre, 2008).

This evaluation was a UK-wide study and six researchers were required to contribute to the collection of observation data. There is always a risk that the use of different researchers to undertake observation of program delivery will produce variations in the practice of observation. Berkel et al., (2011) suggest assessing micro-level behavior s which permit better inter-rater reliability than assessment of more macro-level or qualitative observations. Micro-level behavior s captured in our study include the tone and pace of facilitators, the frequency with which children were encouraged to participate, facilitator familiarity with material and facilitator ability to manage children's behavior (see online Supplement A, for further examples). However, relying solely on a checklist approach will miss capturing all those program components that have a significant impact on outcomes. We used the concept of congruence in assessing and scoring our observations which allowed us to consolidate micro-level behavior s with macro-level behavior s. This was especially important when assessing the quality dimension and, to a lesser degree, the context dimension.

Most of the facilitators observed delivered multiple elements of the program and the specific style of the facilitator/s may have affected IF scoring. We found variability in facilitators' comfort levels in delivering various elements of the program: identifying the required skill-set to deliver all three elements consistently is therefore essential. Additionally, since the same pair of facilitators often deliver all program elements to many schools, a facilitator who lacks confidence with certain aspects of the program can be highly detrimental to the aim of transmitting program messages to its intended beneficiaries.

Working with multiple dimensions when scoring observations leads us to ask the question: is equal weighting between dimensions appropriate? Do all dimensions and their specific remit in IF contribute equally to achieving the desired outcome? How could we better measure the interaction between the various dimensions from both a theoretical and practical perspective?

Our study suggests that implementation fidelity needs to be considered from the inception of a prevention program, starting with the identification of essential components and identifying the extent to which the program may be adapted to certain communities and settings without detracting from delivery of essential components. Implementers should take a key role in managing and negotiating malleable elements of delivery, such as venue suitability and participant readiness. It is important to allow for a certain degree of flexibility in implementation to accommodate diverse contexts and settings, without detracting from the program's key messages.

Who Needs to Understand Implementation Fidelity?

IF should be a familiar concept to all parties involved in developing, delivering and hosting a school-based intervention (Lynas & Hawkins, 2017a). Program developers and delivery teams should ensure that schools are invested in the intervention and ideally have a sense of ownership in respect of the program's key messages which in turn should help children absorb and retain them.

First and foremost, program developers need to assimilate IF into intervention design, taking into consideration the acceptable thresholds to be met for each dimension. This would provide a benchmark with agreed critical components and minimum thresholds to be met to ensure desired outcomes are not compromised.

Our findings suggest that facilitators tasked with delivering programs play a key role and therefore need to understand IF and its importance to programs being delivered as intended. Often those involved in day-to-day delivery are provided with a manual or script but receive limited guidance on fidelity; research on implementers' understanding and operationalization of fidelity is lacking (Cutbush et al., 2017, p. 275). We found coverage of sexual abuse to be low in the workshop element of the program and the process data identified a lack of facilitator confidence in delivering sensitive key messages around sexual abuse. Sexual abuse is one of two topics that make up the critical components of the workshop element and the findings of this study could be used to strengthen facilitator training in these areas. The two-day training package for facilitators may require some attention to developing advanced 'enactment' skills that would build confidence in the materials and prepare facilitators to discuss the topic of sexual abuse in more depth (Lynas & Hawkins, 2017a).

Finally, researchers, implementation teams and evaluators must be aware of IF to ensure outcomes are measured accurately. Taking IF into consideration allows for a fuller understanding of unmet outcomes which may be explained by specific components that are missing from the delivery of the intervention. This could range from content coverage to unsuitable setting and/or unskilled or poorly trained facilitators.

Who Measures Implementation Fidelity?

Results demonstrate that implementation fidelity must play a much greater role in the development of evidence-based interventions for children. The findings underscore the importance of articulating and developing clear scoring protocols using a 'critical component' approach. As Century et al., (2010) highlight, implementation should focus on the components that make up the program. Identifying these components early on and developing them into scoring and guidance protocols (similar to the one we present in Supplement b, available online), would enable program developers and implementers to further focus their training to ensure facilitators have the required skill-set to deliver sensitive interventions within school settings.

Following identification of the critical components within an intervention, they can be well represented within observation schedules and weighted in scoring guidelines to reflect their contribution to achieving outcomes. Training observers prior to evaluating IF will minimise discrepancies in the way observations are recorded and written. Differences between observers in scoring observed sessions created a dilemma for our research team especially when dealing with more qualitative, open-ended comments in the observational records. Our study managed to use these comments constructively, contributing to scores. However, had the observed qualitative comments been recorded in a more defined and systemized way, more insights could have potentially been gained from them.

This study demonstrates the value of integrating implementation fidelity training for researchers, program facilitators and implementers to both assist with monitoring, and to achieve and maintain a standard of quality to mitigate variations in the actual delivery and setting of the program. Measuring fidelity should not be viewed as a one-off activity but should be incorporated within intervention design and monitored on an on-going basis by implementers and facilitators.

Conclusions

The use of a clear structure for assessing implementation fidelity facilitated important findings for this evaluation and learning for program developers. Using a systematic approach to measure the extent of IF allowed us to pinpoint micro-behaviors and conditions when evaluating implementation. We recommend incorporating the principles of IF from the design and development stage of an evidence-based intervention and within the evaluation lifecycle, ensuring they are fully outlined in the evaluation plan. An area that deserves more attention is that of training and professional development for those who deliver prevention programs. This is a much over-looked aspect of implementation fidelity both from a research perspective and program design perspective. Based on these findings, we would suggest that the 'quality of delivery' dimension should be incorporated as an essential rather than an optional aspect of any evaluation.

Assessment of fidelity should also be built into the piloting stage of an intervention. This will allow both developers and implementers the opportunity to ensure that, not only the content but all aspects of delivery, are appropriate and ready for roll-out.

Some uncertainties remain concerning the weighting of scores across the various dimensions - content, quality and context - of IF. Further debates regarding the relative importance of all three dimensions would be valuable for the field of implementation science.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10896-023-00549-z.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Barter, C., Batool, F., Charles, J., Devaney, J., Farrelly, N., Hayes, D., Kurdi, Z., Millar, A., Monks, C., & Foster, H. R., Lorraine Radford, Rhiannon Tudor Edwards, Eira Winrow, and Nicky Stanley. 2022. 'Conducting Large-Scale Mixed-Method Research on Harm and Abuse Prevention with Children under 12: Learning from a UK Feasibility Study'. *Children & Society* online(n/a). doi: https://doi.org/10.1111/chso.12658.

Bauer, M. S., Damschroder, L., Hagedorn, H., Smith, J., & Kilbourne, A. M. (2015). An introduction to implementation science for the Non-Specialist. *BMC Psychology*, *3*(1), 32. https://doi.org/10.1186/s40359-015-0089-9.

Berkel, C., Mauricio, A. M., Schoenfelder, E., & Sandler, I. N. (2011). Putting the Pieces together: An Integrated model of program implementation. *Prevention Science*, *12*(1), 23–33. https://doi.org/10.1007/s11121-010-0186-1.

Bertram, R. M., Karen, A., Blase, & Fixsen, D. L. (2015). Improving programs and outcomes: Implementation frameworks and Organization Change. *Research on Social Work Practice*, *25*(4), 477–487. https://doi.org/10.1177/1049731514537687.

Bruhn, A. L., Shanna, E., Hirsch, & Lloyd, J. W. (2015). Treatment Integrity in school-wide programs: A review of the literature (1993–2012). *The Journal of Primary Prevention*, *36*(5), 335–349. https://doi.org/10.1007/s10935-015-0400-9.

Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., & Balain, S. (2007). A conceptual Framework for implementation Fidelity. *Implementation Science*, *2*(1), 40. https://doi.org/10.1186/1748-5908-2-40.

Century, J., Rudnick, M., & Freeman, C. (2010). A Framework for Measuring Fidelity of implementation: A Foundation for Shared Language and Accumulation of Knowledge. *American Journal of Evaluation*, *31*(2), 199–218. https://doi.org/10.1177/1098214010366173.

Cutbush, S., Gibbs, D., Krieger, K., & Clinton-Sherrod, M., and Shari Miller (2017). Implementers' perspectives on fidelity of implementation: "Teach every single Part" or "Be right with the Curriculum"? *Health Promotion Practice*, *18*(2), 275–282. https://doi.org/10.1177/1524839916672815.

Damschroder, L. J., Aron, D. C., Keith, R. E., Susan, R., Kirsh, J. A., Alexander, & Lowery, J. C. (2009). Fostering implementation of Health Services Research Findings into Practice: A Consolidated Framework for advancing implementation science. *Implementation Science*, *4*(1), 50. https://doi.org/10.1186/1748-5908-4-50.

Dane, A. V., Schneider, B. H., & 'PROGRAM INTEGRITY IN PRIMARY AND EARLY SECONDARY PREVENTION: ARE IMPLEMENTATION EFFECTS OUT OF CONTROL?'. (1998). Clinical Psychology Review18(1):23–45. doi: https://doi.org/10.1016/S0272-7358(97)00043-3.

Durlak, J. A., & Emily, P. D. P. (2008). Implementation matters: A review of Research on the influence of implementation on Program Outcomes and the factors affecting implementation. *American Journal of Community Psychology*, *41*(3–4), 327–350. https://doi.org/10.1007/s10464-008-9165-0.

Fixsen, D. L., & Blase, K. A. (2020). Chapter 3: Active implementation frameworks. *Handbook on implementation science*. Cheltenham, UK: Edward Elgar Publishing.

Fixsen, D. L., Karen, A., Blase, Sandra, F., & Naoom, and Frances Wallace (2009). Core implementation components. *Research on Social Work Practice*, *19*(5), 531–540. https://doi.org/10.1177/1049731509335549.

Fixsen, D., Blase, K., Metz, A., & Melissa Van Dyke (2015). 'Implementation Science'. Pp. 695–702 in *International Encyclopedia of the Social & Behavioral Sciences (Second Edition)*, edited by J. D. Wright. Oxford: Elsevier.

Fixsen, A. A. M., Aijaz, M., & Fixsen, D. L. (2021). Erin Burks, and Marie-Therese Schultes. *Implementation frameworks: An analysis*. The Active Implementation Research Network.

Gearing, R., Edward, N., El-Bassel, A., Ghesquiere, S., Baldwin, J., Gillies, & Evelyn Ngeow (2011). Major ingredients of Fidelity: A review and scientific guide to improving quality of intervention research implementation. *Clinical Psychology Review*, *31*(1), 79–88. https://doi.org/10.1016/j.cpr.2010.09.007.

Griffith, A. K., Hurley, K. D., & Hagaman, J. L. (2009). Treatment Integrity of literacy interventions for students with emotional and/or behavioral Disorders: A review of literature. *Remedial and Special Education*, *30*(4), 245–255. https://doi.org/10.1177/0741932508321013.

Gubbels, J., van der Put, C. E., Geert-Jan, J. M., & Stams, and Mark Assink (2021). Effective components of School-Based Prevention Programs for child abuse: A Meta-Analytic Review. *Clinical Child and Family Psychology Review*, *24*(3), 553–578. https://doi.org/10.1007/s10567-021-00353-5.

Hansen, W. B., Graham, J. W., Wolkenstein, B. H., & Rohrbach, L. A. (1991). Program Integrity as a moderator of Prevention Program Effectiveness: Results for Fifth-Grade students in the adolescent Alcohol Prevention Trial. *Journal of Studies on Alcohol*, *52*(6), 568–579. https://doi.org/10.15288/jsa.1991.52.568

Harn, B., Danielle Parisi, and Stoolmiller, M. (2013). Balancing Fidelity with flexibility and fit: What do we really know about fidelity of implementation in schools? *Exceptional Children*, *79*(2), 181–193. https://doi.org/10.1177/001440291307900205l.

Haynes, A., Brennan, S., Redman, S., Williamson, A., & Phyllis Butow. (2015). The CIPHER team, Gisselle Gallego, and. 'Figuring out Fidelity: A Worked Example of the Methods Used to Identify, Critique and Revise the Essential Elements of a Contextualised Intervention in Health Policy Agencies'. *Implementation Science* 11(1):23. doi: https://doi.org/10.1186/s13012-016-0378-6.

Holloway, J. L., & Pulido, M. L. (2018). Sexual abuse Prevention Concept Knowledge: Low income children are learning but still lagging. *Journal of Child Sexual Abuse*, *27*(6), 642–662. https://doi.org/10.1080/10538712.2018.1496506.

Lundy, L. (2019). A lexicon for Research on International Children's rights in troubled Times. *The International Journal of Children's Rights*, *27*(4), 595–601. https://doi.org/10.1163/15718182-02704013.

Lynas, J., and Russell Hawkins (2017a). Fidelity in School-Based child sexual abuse Prevention Programs: A systematic review. *Child Abuse & Neglect*, *72*, 10–21. https://doi.org/10.1016/j.chiabu.2017.07.003.

Lynas, J., and Russell Hawkins (2017b). Fidelity in School-Based child sexual abuse Prevention Programs: A systematic review. *Child Abuse & Neglect*, *72*, 10–21. https://doi.org/10.1016/j.chiabu.2017.07.003.

Rojas-Andrade, R., & Loreto Leiva, B. (2019). Is implementation Fidelity important? A systematic review on School-Based Mental Health Programs. *Contemporary School Psychology*, *23*(4), 339–350. https://doi.org/10.1007/s40688-018-0175-0.

Sanetti, L. M., Hagermoser, L. M., Dobey, & Gritter, K. L. (2012). Treatment Integrity of Interventions with children in the Journal of positive behavior interventions from 1999 to 2009. *Journal of Positive Behavior Interventions*, *14*(1), 29–46. https://doi.org/10.1177/1098300711405853.

Sarno, J., Owen, A. R., Lyon, N. E., Brandt, C. M., Warner, E., Nadeem, C., Spiel, & Mary Wagner. (2014). Implementation science in School Mental Health: Key constructs in a developing Research Agenda. *School Mental Health*, *6*(2), 99–111. https://doi.org/10.1007/s12310-013-9115-3.

Stanley, N., Devaney, J., Kurdi, Z., Ozdemir, U., Barter, C., Monks, C., Edwards, R. T., Batool, F., Charles, J., Farrelly, N., Hayes, D., Millar, A., Thompson, T., Winrow, E., & Lorraine Radford. (2023). and. 'What Makes for Effectiveness When Starting Early – Learning from an Integrated School-Based Violence and Abuse Prevention Programme for Children Under 12'. *Child Abuse & Neglect* (139).

Stanley, N., Barter, C., Batool, F., Charles, J., Devaney, J., Edwards, R. T., Farrelly, N., Hayes, D., & Kasperkiewicz, D., Berni Kelly, Zain Kurdi, Annemarie Millar, Claire Monks, Ugur Ozdemir, Lorraine Radford, Trevor Thompson, and Eira Winrow. 2021. *Evaluation of the NSPCC Speak out Stay Safe Programme: Final Report*.

NSPCC. n.d (September 2022). 'Speak out Stay Safe Programme'. *NSPCC Learning*. Retrieved 12 (https://learning.nspcc.org.uk/services/speak-out-stay-safe/).