Case Study 1: An Evidence-Based Practice Review Report

How effective are adapted versions of the family-based intervention, SFP:10-14, for reducing problem behaviours in children aged ten to fourteen years?

Summary

The Strengthening Families Programme: Parents and Youth 10-14 (SFP:10-14) is a family-based intervention developed in the United States (US, Molgaard & Spoth, 2001). It uses an interactive skills-building curriculum to promote effective parenting; positive communication; and problem solving strategies to reduce substance use and problem behaviours in children aged ten to fourteen years (Molgaard & Spoth, 2001). Promising research has emerged from the US indicating SFP:10-14 is an effective intervention, with medium to large effect sizes reported (Kumfer, Whiteside, Greene, & Allen, 2010). In response, several studies using culturally adapted versions of SFP:10-14 have been replicated worldwide, with inconsistent reports for its efficacy. Subsequent reviews and meta-analyses of SFP:10-14 have been published, however these vary by participant age and include outcomes from measures of substance use only. Furthermore, they fail to isolate adapted versions of SFP:10-14 from the original version, such that its effectiveness once adapted can reliably be inferred. Thus, this review is considered to be the first.

Five studies were included in the review; following a database search and systematic matching to a pre-determined inclusion criteria. These studies were then reviewed using Harden and Gough's (2012) Weight of Evidence Framework, and their methodological quality appraised using an adapted version of the Kratochwill group-based design coding protocol (Kratchowill, 2003). Adapted versions of SFP:10-14 were found not to be effective in reducing problem behaviours when replicated and implemented outside the general population in the US. Effect size calculations support conclusions drawn, with small effect sizes found for non-significant results reported in four studies, and a small effect size found for one study reporting positive, significant results. Limitations of this review are provided, and recommendations for future research are discussed.

Introduction

Extant literature supports the effective functioning of the family as an important determinant of delinquency prevention (Case & Haines, 2005). In particular, family variables such as: lack of parental supervision, parental rejection, and erratic and harsh discipline, have been reported as significant predictors of anti-social behaviours in adolescents (Hawkins, Catalano, & Miller, 1992). Historically, intervening in the private domain of the family was considered intrusive, such that intentions of supporting families was considered secondary to their right for privacy (Graham & Bennett, 1995. As cited in Knepper, 2007). However, in the UK, with the advent of the 'parenting order' (Crime and Disorder Act, 1998); accountability and responsibility was placed upon parents to safeguard effective family function to minimise maladapted functioning in

young people (Case & Haines, 2005). Despite this important legislative change, and the apparent convergence in the literature to support the need for family-based interventions (Ross, Duckworth, Smith, Wyness, & Schoon, 2010), there is a paucity of evidence to support the effectiveness of any one parenting intervention over another. However, one intervention, the Strengthening Families Programme: For Parents and Youth 10-14 (SFP:10-14, Molgaard & Spoth, 2001) has widely been considered to show promising outcomes (Allen, Coombes, & Foxcroft, 2007).

What is SFP:10-14?

The SFP:10-14 programme is a time sensitive revision of the 14-week Iowa Strengthening Families Programme (ISFP; Spoth, Redmond & Shin, 2000). It is a highly structured and standardised, video-based programme; designed to facilitate parents and children learning together. It has an overall objective of improving parenting practices and helping parents and children develop skills known to reduce risk and improve protective factors in children (Molgaard & Spoth, 2001). For example, during the first hour, parents are taught skills, such as: positive attention and rewards, as well as appropriate disciplinary practice. Meanwhile, the children are taught complimentary skills, such as: anger management, social skills, positive communication, and compliance with parental rules. By providing parallel content and shared family activities, learning is argued to be re-enforced more effectively than if children and parent activities are carried out in isolation of each other (Allen et al., 2007). The SFP:10-14 programme runs for two hours per week, over the course of seven weeks; further supplemented by four voluntary booster sessions. It is

delivered by three group leaders at approved sites, including: schools, social service agencies, churches, or community centres. Importantly, group leaders are specially trained, and are responsible for leading and facilitating the intervention for the same families each week.

SFP:10-14 and Psychological Theory

SFP:10-14 is underpinned by the biopsychosocial vulnerability model (Kumfer, Trunnel, & Whiteside, 1990). This model provides a framework to help illuminate family risk variables, such as conflict and financial stress, and improve family coping skills, such as conflict resolution, communication skills, and social support; all of which are considered to interact with a child's community and peer related variables to influence their adjustment outcomes (Molgaard, Spoth, & Redmond, 2000). Embedded within this framework is a developmental perspective. This assumes early adolescence is a tumultuous period, demarcated by an increase in disagreement and conflict between young people and their families (Baer, 1999). It is also considered to be a stage when experimentation with risky behaviours typically begins; exacerbated by elevated levels of stress as adolescents navigate the emotional upheaval of transitioning from primary to secondary school (Spoth, Guyll, Chao, & Molgaard, 2003). Therefore, interventions like SFP:10-14 that include the family, and target early adolescence, are argued to be effective in reducing the risk of more pervasive, longer term, problem behaviours (Rothbaum & Weisz, 1994).

SFP:10-14 Adaptions and the Empirical Evidence

A Cochrane systematic literature review (Foxcroft, Ireland, Lister-Sharp, Lowe, & Breen, 2002) and a recent meta-analysis (de Vicente et al., 2017), which evaluated interventions for the primary prevention of substance use, do suggest the Strengthening Families Programme (across various ages) is an effective and promising programme. Furthermore, meta-analyses of school based (Foxcroft & Tertsvadze, 2011) and non-school based interventions for drug use (Gates, McCambridge, Smith, & Foxcroft, 2006), report SFP:10-14 is twice as effective as any programme at preventing alcohol and drug use. However, to date, there has not been a systematic literature review to investigate antisocial and problem behaviours as secondary outcomes of SFP:10-14. This is a rather surprising finding, and considered herein to be an unfortunate oversight.

Disruptive behaviours are targeted in SFP:10-14 due to their likely mediating role in adolescent substance use (Molgaard & Spoth, 2003). Furthermore, moderate effect sizes have been reported in US studies that measured outcomes of disruptive behaviour (Spoth, Randall, Trudeau, & Shin, 2008; Spoth, Trudeau, Redmond, & Shin, 2016). Therefore, published reviews, which investigate outcomes of substance use only, do seem to overlook the trajectory for young people with problem behaviours and its impact on other life domains. For example, problem behaviour in adolescence has been linked to poorer outcomes in school engagement, academic attainment (Goodman & Gregg, 2010), employment prospects, as well as health (Allen et al., 2007).

Thus, reviewing the utility of SFP:10-14 for reducing problem behaviours, independently of any effect on substance use alone, is well supported.

In addition, this review will evaluate the effectiveness of adapted versions of SFP:10-14 in attempt to elucidate possible differences between reviews which incorporate all studies of SFP:10-14. It is widely reported, that promising evidence from research using high-quality, randomised controlled trials in the US (Allen et al., 2007), provided the impetus for the replication of SFP:10-14, and several studies using adapted versions prevailed in Panama (Mejia, Ulph, & Calam, 2016); the US Pacific Northwest (Roulette, Hill, Diversi, & Overath, 2017); Spain (Perez et al., 2012); Italy (Ortega, Latina, & Ciarano, 2012); and the UK (Coombes, Allen, Marsh, & Foxcroft, 2009). Individually, these studies aimed to reduce the cultural and linguistic distance between SFP:10-14 resources and the target group, whilst remaining faithful to the original version. However, confounding limitations, such as: exploratory designs; small sample sizes; no control groups; questionable reliability and validity of measurement tools; and selection bias; marred the generalisability of their results (Coombes, Allen, & Foxcroft, 2012). Furthermore, concerns were raised for substantive adaptions, which resulted in programmes no longer resembling the original, such that effects could not be reliably attributed to SFP:10-14 (Gorman, 2017). Therefore, it seems prudent to review adapted versions of SFP:10-14 such that its effectiveness for reducing problem behaviours in children, can be fully understood.

SFP:10-14 and Educational Psychology

The implications of evidence-based interventions, which are effective in reducing problem behaviour in adolescents, is argued herein to be especially relevant to Educational Psychology. For example, problem behaviour has been associated with lower school achievement in adolescents (Goodman & Gregg, 2010) and leads to an increased risk of school exclusion (Skinner, Kindermann, & Furrer, 2008). Furthermore, early adolescence, and especially the transition from primary to secondary school, is widely accepted as a challenging time for young people; but especially for those with problem behaviours (Molgaard & Spoth, 2001). Thus, SFP:10-14 could offer Educational Psychologists an alternative intervention to provide support for parents and young people. It could help to improve child outcomes in school, as well as mediating for other risky behaviours, which may otherwise interfere with their positive adjustment through adolescence.

To that end, this review will seek to answer the following question:

How effective are adapted versions of the family-based intervention, SFP: 10-14, in reducing problem behaviours in children aged ten to fourteen years?

Critical Review of the Evidence Base

Literature Search

A comprehensive literature search of the databases Web of Science, PsychINFO and ERIC was conducted between 3rd and 4th of January 2019. A Title and Abstract search included the following terms: see Table 1.

Table 1.

Data Base Search and Results

Data Base	Search Terms Used	Total Results
Web of Science	**TOPIC** ("Strengthening Families" OR "Iowa Strengthening Families" OR "Strengthening Families Program*" OR "SFP 10-14") *AND* **TOPIC:** ("Problem Behavio*r" OR "Disruptive Behavio*r" OR "Challenging Behavio*r" OR "Conduct" OR "Delinquen*" OR "Behavio*r"	59
ERIC (EBSCO)	"SFP 10-14" OR "Strengthening Families" OR "IOWA Strengthening Families"	60
PsychINFO	"Strengthening Families" OR "Iowa Strengthening Families") *AND* "Behavio*r	61

*Denotes wildcard

All together, these searches generated 180 results. Once duplicates were removed (n=25 duplicates), the remaining articles (n=155) were screened by title (n=88 excluded by title) and abstract (n=52 excluded by abstract) against

the inclusion and exclusion criteria detailed in Table 2. Fifteen articles were left for full text screening, ten of which were excluded for failing to meet the inclusion criteria. These articles are fully described in Appendix A. Ancestral searches of articles selected for inclusion revealed no new articles. Authors were emailed directly to solicit missing data or to request more information to support a report too brief so as to be otherwise excluded (Coombes et al., 2012; and Foxcroft et al., 2016). Supplementary Data was also retrieved from EURPUB for articles missing the information required to make judgments for inclusion (Baldus et al., 2016; Coombes et al., 2012; and Foxcroft et al., 2016). Five articles were selected for in-depth analysis in this review, these are summarised in Table 3 and are fully described in Appendix B. Figure 1 depicts the study selection process.

Table 2.

Inclusion and Exclusion Criteria

Criteria	Inclusion Criteria	Exclusion Criteria	Rationale
1. Type of Publication	Study must be published in a Peer Reviewed Journal	Study was not published in a peer reviewed journal. (e.g. Book Chapter, Review, Doctoral Thesis or Dissertation	Study must meet criteria for academic rigour (reliability)
2. Intervention	Study must be an approved adaption of the Strengthening Families Programme:10-14 (SFP:10-14)	Study is an adaption of Strengthening Families Programme (e.g. Unrecognised as approved by the original intervention authors) Study is another age version of SFP: (3-6), (6-11) or (12-16) Study included Strengthening Families Programme as an adjunct to another Intervention (e.g. 'Mindfulness'). Study is an alternative (e.g. Iowa Strengthening Families Programme ISFP or Strengthening Families Strengthening Communities)	Required to critically review the effectiveness of SFP:10-14.
3. Study Design	Study includes primary empirical data derived from Randomised Controlled Trials or Quasi-Experimental Design	Study does not include primary empirical data (e.g. book chapter or review) or empirical data (e.g. case study) or does not have a control group	RCTs and Quasi-experimental design are considered to be superior in reliability and validity when compared with non-experimental designs
4. Participants	Participants must include parents/carers (mother and/or father; female and/or	Adult participants are teachers or other community adults.	Required to critically appraise the effectiveness of a parent intervention for improving child outcomes.

Criteria	Inclusion Criteria	Exclusion Criteria	Rationale
	male) and children aged 10 years to 14 years.	Child participants are older or younger than 10yrs -14yrs	Required for evaluating the specific age version for SFP:10-14
5. Outcome Measures	Study reports child outcomes Study reports child externalising and/or problem behaviour outcomes Study reports post intervention outcomes taken at the earliest point	Study does not report child outcomes (e.g. parent outcomes) Study reports other child outcomes (e.g. alcohol and drug use, problem solving skills) Study only reports follow up data (e.g. 4 year post initial analysis)	Required to critically appraise the effectiveness of SFP 10-14 on child externalising or problem behaviours
6. Language	Study must be in English	Study not in English	No resources for reliable translation

Figure 1.

Flow Diagram of the Study Screening Process

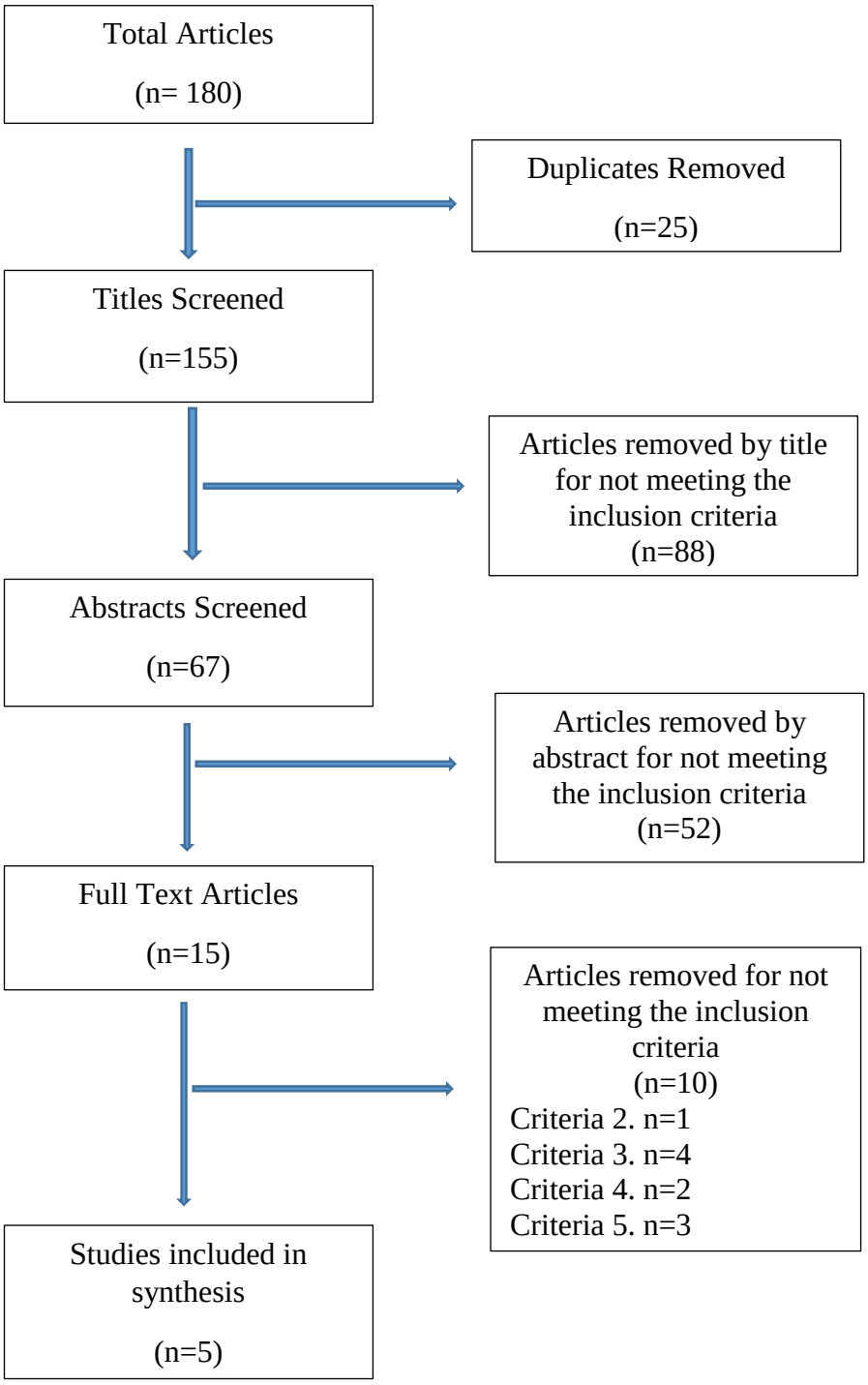

Table 3

Summary of Studies Included in Review

Baldus, C., Thomsen, M., Sack, P. M., Bröning, S., Arnaud, N., Daubmann, A., & Thomasius, R. (2016). Evaluation of a German version of the Strengthening Families Programme 10-14: A randomised controlled trial. *European Journal of Public Health*, *26*(6), 953–959. https://doi.org/10.1093/eurpub/ckw082
Coombes, L., Allen, D. M., & Foxcroft, D. (2012). An exploratory pilot study of the Strengthening Families Programme 1014 (UK). *Drugs: Education, Prevention and Policy*, *19*(5), 387–396. https://doi.org/10.3109/09687637.2012.658889
Foxcroft, D. R., Callen, H., Davies, E. L., & Okulicz-Kozaryn, K. (2016). Effectiveness of the strengthening families programme 10-14 in Poland: Cluster randomized controlled trial. *European Journal of Public Health*, *27*(3), 494–500. https://doi.org/10.1093/eurpub/ckw195
Spoth, R., Guyll, M., Chao, W., & Molgaard, V. (2003). Exploratory Study of a Preventive Intervention With General Population African American Families. *The Journal of Early Adolescence*, *23*(4), 435–468. https://doi.org/http://dx.doi.org/10.1177/0272431603258348
Skärstrand, E., Sundell, K., & Andréasson, S. (2014). Evaluation of a Swedish version of the Strengthening Families Programme. *European Journal of Public Health*, *24*(4), 578–584. https://doi.org/10.1093/eurpub/ckt146

Weight of Evidence

To critically analyse the quality and relevance of evidence for each of the five studies, the Harden and Gough's (2012) weight of evidence (WoE) framework was applied. This provides for judgements to be made across three dimensions: methodological quality (WoE A); methodological relevance (WoE B); and relevance to the research question (WoE C). Scores for all three dimensions were equally weighted and averaged to provide an overall weight of evidence – WoE D (See Appendix C for a full description and breakdown).

All included studies were coded using an adapted version of the Kratochwill Group-Based Design Protocol (2003). Modifications accounted for the specificity of the research question and certain sections were either considered

irrelevant to the study and were removed, or especially relevant and subsequently elaborated upon for later discussion. Appendix D provides justification for omitting or elaborating some sections. Appendix E provides all individual study protocols.

WoE D was calculated using Upper and Lower markers split into terciles and used to reflect the range; scored and rated as follows: Low = 1 to 1.6 Medium = 1.7 to 2.3; and High = 2.4 to 3.0. Table 4 provides a summary of the WoE for the selected studies, including their mean score (in brackets).

Table 4.

Weight of Evidence Evaluated for Each Study

Study	WoE A: Quality of Methodology	WoE B: Relevance of Methodology	WoE C: Relevance of evidence to the review question	WoE D: Overall weight of Evidence
Baldus et al., (2016)	High (3)	High (3)	High (3)	High (3)
Coombes et al., (2012)	High (3)	Medium (2)	Medium (2)	Medium (2.3)
Foxcroft et al., (2016)	High (3)	High (3)	Medium (2)	High (2.7)
Skarstrand et al., (2014)	High (3)	Medium (2)	Medium (2)	Medium (2)
Spoth et al., (2003)	Medium (2)	Medium (2)	High (3)	High (2.3)

Participants

The number of participants in the five studies ranged from 69-614. In accordance with the inclusion criteria, all participants were aged between 10-14 years (M=11.68). Girls (n=804) were slightly over-represented (boys:

n=753), however no studies reported sex or gender effects relevant to the review question.

The studies were conducted in different countries including the US (n=1), Sweden (n=1), UK (n=1), Poland (n=1), and Germany (n=1). Whilst this review did not exclude by geographic region, it is rather interesting to note that all included studies are members of the OECD. This raises the issue of the cost implications that pervade RCTs, and although it is not within the scope of this review to appraise critically, it is considered to be an important observation.

One study included participants assessed as 'at risk' at baseline (Baldus et al., 2016), and two studies included participants from low SES backgrounds (Foxcroft et al., 2016; Spoth et al., 2013). For these two studies, low SES was reported as a risk factor for child problem behaviours, and were considered to better represent the intervention's target population. Subsequently, all three were rated higher for WoE C than those studies which included participants from the general population (Coombes et al., 2012; Skarstrand et al., 2013).

Sampling method was reported in all five studies. Three studies used a single stage cluster sample method (Coombes et al., 2012; Foxcroft et al., 2016; Skarstrand et al., 2013), and two studies used a two-stage cluster method by randomly sampling from a larger pool of African Americans (Spoth et al., 2003) or from low SES communities (Baldus et al., 2016). Cluster sampling is widely considered to be more efficient when a study takes place over a large geographical region, as was the case for all studies included for this review. Whilst cluster sampling combines the benefits of random and stratified sampling by allowing the accumulation of large samples, they are at an

increased risk of selection bias (Hahn, Puffer, Torgerson, & Watson, 2005). For example, if the chosen clusters are not representative of the population, it would impact the confidence needed to make inferences about the target population. However, all studies satisfactorily report steps taken to confirm representation of the target population and were subsequently equally rated for WoE A.

Power analysis was calculated and reported in the original version of SFP:10-14 (Spoth et al., 1999), wherein 580 participants were deemed necessary to allow a detection of a moderate effect size of 0.30 (95% power at the $p<0.05$ level of significance). All five studies sought to replicate SFP:10-14 and referred to this power analysis to inform their own sample size expectations. Two studies were sufficiently powered (Foxcroft et al., 2016; Skarstrand et al., 2013), however three studies had insufficient sample sizes (Baldus et al., 2016; Coombes et al., 2012: Spoth et al., 2003), consequently reducing the confidence with which any findings or effect sizes could be considered.

Attrition of participants provided the impetus for the development of SPF:10-14. Therefore, low attrition was considered to be especially relevant, indeed; one study sought to mitigate for 'drop out' by randomly assigning to the intervention and control group by a ratio of 2:1 (Foxcroft et al., 2016). All studies reported attrition rates and conducted statistical analyses to correct for missing data. Although no differences are reported, authors do discuss attrition as a possible limitation to the generalisability of their studies. Subsequently, studies were coded higher when below 20% attrition, even

where intention-to-treat analyses were applied and reported as insignificant.

Research Design

Four of the five studies included for this review used a Randomised Controlled Trial design (RCT). RCTs increase the internal validity of a study by using control groups to allow for distinctions to be made between effects of the intervention and other confounding extraneous variables which may originate within participant groups. Subsequently, RCTs were rated higher than other research designs. Furthermore, whilst all four studies stratified differing variables, including socio-demographic variables of age and gender, as well as other additional variables, such as: highest education of the responding parent; ethnicity of child, geographic social load; and school grade, they all reported no significant differences between groups at baseline, and were therefore equally weighted for WoE B.

One study however, was a quasi-experimental design without randomisation (Coombes et al., 2012). Whilst Coombes and colleagues (2012) initially designed their study as a cluster RCT, the desire of some participants to be included in the intervention group resulted in them changing their research design rather than risk abandoning the study altogether. Although they report no significant difference between groups at baseline, it was scored lower for methodological relevance compared to the other four studies (WoE B: Table 3).

Study designs which include scope to evaluate maintenance of effects, were considered herein to be superior to those which do not. Therefore, the quality of methodology (WoE A) included a rating for follow up data. Three studies

included a follow-up (Baldus et al., 2016; Foxcroft et al., 2016; Skarstrand et al., 2013) and were scored higher than those which reported immediate effects only (Coombes et al., 2012; Spoth et al., 2003).

Measures

Standardised measures were used by all five studies to report pre-and post-intervention behaviours (Full summary in Appendix B). However, only two studies used the measures as employed in the original SPF:10-14 (Coombes et al., 2012; Spoth et al., 2003). Although consideration was given to rating these studies higher, it was ultimately decided that, where alternative measures were valid and had high internal consistency, ratings would be commensurate with studies using the original measures. This reasoning did not follow for rating multi-source measures. That is, studies were rated higher for methodological relevance when they included child self-report and parent report (Baldus et al.,2016; Coombes et al., 2012), compared to those which included child self-report only (Foxcroft et al., 2016; Skarstrand et al., 2013; Spoth et al., 2003). Having multi-sources of data was considered to help minimise bias, whereupon a child's self-report of problem behaviour may be understated because they don't necessarily suffer from it. Alternatively, a parent may respond differently as they are more likely to suffer from the problem behaviour. Whilst modest cross-informant agreement is widely agreed to be a robust phenomena in child and adolescent mental health research (De Los Reyes et al., 2015), the present author appraised it to be more valuable to have both respondent's data to help answer the research question.

Intervention

Details for how each study adapted and implemented SPF:10-14 are summarised in Table 5. Since fidelity to the original SPF:10-14 is fundamental to answering the review question, replication characteristics were considered especially important and were elaborated upon for consideration in judgements using WoE A protocol.

All studies were approved by the original developers of SFP:10-14, however, one study was rated lower for WoE B because it made adaptions reasonably considered to be substantial enough to impact the extent to which inferences could be made from their results (Skarstrand et al., 2013). In addition, closer examination of replication features were carried out during protocol coding. This indicated that Foxcroft et al., (2016) failed to report facilitator characteristics and intervention site locations, and Spoth et al., (2003) combined the 6th and 7th intervention sessions; both arguably increasing the risk of introducing confounding variables. However, the two studies go on to discuss the rigour used to follow programme protocol. Therefore, it seems reasonable to assume that appropriate training was most likely carried out by Foxcroft and colleagues (2016), and that the combining of two final sessions was a surface level rather than a significant revision made by Spoth and colleagues (2003). To that end, with the exception of Skarstrand and colleagues (2013), four studies are considered to have provided robust evidence to support their adherence to the original SPF:10-14, and that all adaptions, whilst pervasive, are considered desirable accommodations, which are more likely to enable real-world applications.

Table 5.

Intervention Summary and Adaptions

Study	Programme Name	Adaption Process and approval	Trainer	Duration	Session length	Booster	Group Size	Setting
Baldus et al., 2016	SPF:10-14 German Version	Cultural adaptions approved in stepped phase for RCT (Stolle et al., 2011) Closely aligned with original SPF:10-14 US version – Equivalence of manual and video material, programme delivery in parallel parent and child sessions with joint family activities and meals	Three trained staff members	7 weekly sessions	2 hours per sessions	Optional 4 booster	8-12 Families	Local youth welfare and addiction aid organisation
Coombes et al., 2012	SPF:10-14 UK Version	Cultural adaptions approved in stepped phase for RCT (Coombes & Foxcroft, 2007) and was based on the Medical Research Council (MRC) framework for development and evaluation of RCT's for complex interventions Closely aligned with original SPF:10-14 US version - Equivalence of manual and video material, programme delivery in parallel parent and child sessions with joint family activities and meals	Three five-person teams Three day training programme	7 weekly sessions	2 hours per session	No booster	8-12 Families	Two Schools and one community centre
Foxcroft et al., 2016	SPF:10-14 Polish Version	Cultural adaptions approved in stepped phase for RCT design (Okulicz-Kozaryn et al., 2012)	Not reported	7 weekly sessions	2 hours per session	Optional 4 booster	15 families	Not reported

Study	Programme Name	Adaption Process and approval	Trainer	Duration	Session length	Booster	Group Size	Setting
		Closely aligned with original SPF:10-14 US version - Equivalence of manual and video material, programme delivery in parallel parent and child sessions with joint family activities and meals (supplementary data: Okulica-Kozaryn et al., 2012)						
Skarstrand et al., 2013	SPF:10-14 Swedish Version All schools from	Cultural adaptions approved by the programmes first author, Dr Virginia Molgaard. Original version video and manuals almost identical but for language translations. Distinct differences from the original format including six separately held sessions with one joint family session (part one). Optional booster session were turned into a regular part (part two) with one added session. Some family sessions were omitted Additional materials in part two (substance use)	14 Leaders and 20 teachers. Trained by certified SFP:10-14 trainers	14 weekly session	2-3 hours per session	Booster included as regular part of intervention	Not reported	Schools
Spoth et al.,	SPF:10-14 for African American families	Cultural adaptions approved by the programmes first author, Dr Virginia Molgaard. Closely aligned with original SPF:10-14 US version - Equivalence of manual and video material, programme delivery in parallel	Facilitators and observers received 2 days certified training	6 weekly sessions 7th session subsumed into the final of the 6th session	2 hours per session	Optional 4 booster	5-10 families	Schools and community centres.

Study	Programme Name	Adaption Process and approval	Trainer	Duration	Session length	Booster	Group Size	Setting
		parent and child sessions with joint family activities and meals. Modifications consisted of African American narrators and actors for video and other materials. Artwork considered appropriate to the target population Youth sessions were delivered with consideration given to an active learning style considered to be preferred by African American adolescents (Hale-Besnon, 1982).						

Outcomes and Effect Sizes

Outcomes and effect sizes for all five studies are summarised in Table 6. For this review, Cohens *d* (1988) was applied as the benchmark for appraising effect size. This indicates a small effect (*d*=0.2), a medium effect (*d*=0.5) and a large effect (*d*=0.8). Only one study reported a significant effect of SFP:10-14 on problem behaviours (Spoth et al., 2003), however the effect size was so small that it is reasonably described as negligible (Cohen's *d* = 0.03). This effect size indicates that, while the children receiving the intervention reported a reduction in aggressive and hostile behaviours, the difference between the groups is too small to reliably infer the outcome is a result of the intervention (Thompson, 2007). This study provides an example for the importance of carrying out effect size statistics to avoid spurious relationships underpinning misleading information when significance alone is reported.

Coombes et al., (2012) reported non-significant findings for the effectiveness of SPF:10-14, however they did not provide statistics for effect size calculations to be carried out. The remaining three studies all reported non-significant findings and calculations revealed extremely small (negligible) effect sizes (Baldus et al., 2016 Cohen's *d* = -0.07 and *d* = 0.02; Skarstrand et al., 2014 Cohen's *d* =-0.06). Notably, Foxcroft et al., (2016) reported credible intervals to support Bayesian regression models of analysis. Deines (2014) suggests that using Bayes factor and Credible Intervals allows for the interpretation of non-significant results, helping to distinguish between whether results support a null hypothesis or if data are simply insensitive. Accordingly, attempts were not made to convert to a frequentist approach for

calculating effect size (Cohen’s d), but to instead accept and interpret credible intervals as a reliable reflection of the intervention effect.

Table 6

Effect Sizes of Included Studies

Study	Participant Sample	Outcome	Significant (Between Gps)	Effect Size	Qualitative Description	Overall WoE D
Baldus et al., (2016)	n= 292 (intervention n=147)	T1 to T2 (8 weeks) Child Report: Reduction in anti-social behaviour – RAASI	Not significant (M=0.20, 95% CI = -0.86 to .46, p = 0.29)	Cohens d^a = -0.07	Negligible effect	
		Parent Report: Improvement in externalising problem behaviour	No significant differences between groups (Adjusted M=0.03, 95% CI = -35 to 0.41, p = 0.7)	Cohens d =0.02	Negligible effect	High
Coombes et al., (2012)	n=69 (intervention n=34)	T1 to T2 (Last session) Child self-report: Reduction in Aggressive and destructive behaviour Parent Report: Reduction in destructive and aggressive behaviour	No statistics available – Author reports no significant effect of SFP 10-14 for control group differences for any variables examined		N/A	Medium
Foxcroft et al., (2016)	n=614 (intervention n=367)	T1 to T2 (12 months) Child Self-report: Reduction in Aggressive and destructive behaviour	No significant effect. Multiple imputed propensity score matched for complete case (CC^c) using posterior mean ratios CC = 0.00	CI^b = -0.18 – 0.11	Small effect	

Study	Participant Sample	Outcome	Significant (Between Gps)	Effect Size	Qualitative Description	Overall WoE D
		Child Self-report: Improvement in externalising subscale of SDQ	No significant effect Multiple imputed propensity score matched for complete case using posterior mean ratios CC = -0.10	CI = -0.23 - -0.03	Small to Negligible effect	High
Skarstrand et al., (2014)	n=587 (intervention n=371)	T1-T2 (12 months) Child Self-report: Reduction in norm-breaking behaviours	No Significant effect. OR** = 1.01 95% confidence interval = 0.59-1.71	Cohens *d* =-0.06	Negligible effect	Medium
Spoth et al., (2003)	n=85 (intervention n=34)	T1-T3 (8 weeks) Child self-report: Reduction in Aggressive and destructive behaviour	Significant $p < .005$	Cohen's $d = 0.03$	Negligible effect	High

[a] *d* = Describes effect size (Cohen, 1988)

[b] CI = Confidence Intervals

[c] CC = Complete Case. Describes analysis type when cases with missing values are omitted from data analysis. Foxcroft et al., 2016)

Discussion

This systematic review aimed to evaluate the effectiveness of adapted versions of SPF:10-14 for reducing problem behaviours in children and young people. Accordingly, it is was found that there is little evidence to support adapted versions SFP:10-14 as an effective intervention when replicated and implemented outside of the general population in the US. Conclusions are underpinned by the combination of non-significant results, further supported by calculations for effect size, as well as the synthesis of other weight of evidence criteria from which conclusions can be inferred. This includes: study methodological quality, relevance and specificity to the research question.

The lack of evidence supporting SPF:10-14 is somewhat surprising and entirely inconsistent with trial results from the US. One explanation could be the internal validity of the programme intervention; described as a Type III error. That is, replicated studies fail to implement an intervention as it was originally intended (Israel et al., 2005). Indeed, Skarstrand and colleagues (2013) have been criticised for altering substantive mediators of change in their programme. Whilst they defend their revisions as a practical necessity, it is a significant limitation. Despite this, the remaining four studies were evaluated as providing sufficient evidence to indicate their adaptions were reasonable, and that their adherence to the original SFP:10-14 programme was robust. Consequently, study fidelity and replication is not considered to explain the lack of effect found.

Interestingly, due to the heterogeneity between the studies for the timing of post-test data collection, a relative effect could reasonably have been anticipated. However, regardless of whether the post-test occurred at the last session, 4 weeks or 12 months after intervention completion, no significant effect was reported. Moreover, the one study to report a reduction in problem behaviours had an effect size so small that confidence in the results is largely diminished. This, together with the strong collective weight of evidence supporting intervention fidelity, supports the conclusions made in this review.

Limitations

This review sought to evaluate the effectiveness of replicated studies of SFP:10-14 in isolation of any real comparison with the original version. Arguably, without having applied the same critical review criteria to original studies, one cannot be sure that the findings reported are robust so as to provide for a fair comparison. Indeed, Gorman (2017) in a review of original SFP:10-14 trials, suggested that the reported effects for substance use were actually "*chance findings emerging from flexible data analysis and selective reporting"* (p.29). Therefore, it is unclear whether the lack of effectiveness in replicated studies for reducing problem behaviours is due to confounds introduced through revisions, such as extraneous variables like cultural differences; or whether findings reported simply reflect the true effect of the original SFP:10-14.

Another limitation is the risk of bias associated with statistical analysis using Intention-to-Treat (ITT). For example, ITT is considered to provide conservative corrections for missing responses by maintaining a prognostic

balance produced from randomisation (Gupta, 2011). Certainly, it was the approach accepted in the present review, whereupon, imputed data was reported rather than data before ITT analysis was carried out. This was necessary to ensure large enough sample sizes to ensure sufficient power. However, Gupta (2011) cites reliable evidence to argue that the efficacy of an intervention is in doubt when a participant who did not receive the intervention is included like a participant who did. Furthermore, by pooling missing data, the risk of introducing heterogeneity is increased when non-compliant and compliant subjects are mixed together. Consequently, caution is advised when interpreting significant results and effect sizes, however, it seems reasonable to assume caution should also apply to non-significant results, such as those that pervade this review.

Finally, this review had a relatively small number of studies available for evaluation, furthermore, the studies included are from different countries. This heterogeneity makes considerations for any contextual influences challenging. This is exacerbated by the lack of other qualitative data, which could have provided a nuanced perspective of participants, such that explanations for the lack of effectiveness may well have been better understood.

Recommendations

Recommendations for future research include an investigation of possible cultural explanations for why SFP:10-14 seems to fail to replicate positive findings in the US. This should include a rigorous critical review of the original findings carried out by impartial researchers not associated with the larger research team in the US. It could also include other methodological designs

to provide data, which quantitative approaches alone may not provide. It may also be helpful to explore differences between other age versions of SFP replicated outside of the US to help contextualise effects to child and adolescent developmental factors.

These recommendations are considered to be especially important in light of Educational Psychology practice in the UK, which has evolved to include a present emphasis on provisions and recommendations supported by a strong evidence base (Frederickson, 2002). In particular, this review highlights the problematic nature of transporting evidenced based programmes from one country to another. That is, to assume intervention effectiveness will prevail in the UK because of success in the US, fails to account for possible social and cultural differences which may act as extraneous variables impeding the effectiveness of SFP: 10-14. Therefore, EP's should be cautious and maintain a critical perspective when considering making recommendations for interventions where the evidence base does not include UK samples.

References

Allen, D., Coombes, L., & Foxcroft, D. R. (2007). Cultural accommodation of the Strengthening Families Programme 10 – 14 : UK Phase I study, *22*(4), 547–560. https://doi.org/10.1093/her/cyl122

Baer, J. (1999). The effects of family structure and SES on family processes in early adolescence. *Journal of Adolescence,* 22, 341-354.

Case, S., & Haines, K. (2015). Positive Promotion : Reframing the Prevention Debate. https://doi.org/10.1177/1473225414563154

Coombes, L., Allen, D., Marsh, M., & Foxcroft, D. (2009). The Strengthening Families Programme (SFP) 10-14 and Substance Misuse in Barnsley: The Perspectives of Facilitators and Families. *CHILD ABUSE REVIEW*, *18*(1), 41–59. https://doi.org/10.1002/car.1055

Coombes, L., Allen, D. M., & Foxcroft, D. (2012). An exploratory pilot study of the Strengthening Families Programme 10-14 (UK). *DRUGS-EDUCATION PREVENTION AND POLICY*, *19*(5), 387–396. https://doi.org/10.3109/09687637.2012.658889

Crime and Disorder Act. (1998). UK. Available at: https://www.legislation.gov.uk/ukpga/1998/37/pdfs/ukpga_19980037_en.pdf.

De Los Reyes, A., Augenstein, T. M., Wang, M., Thomas, S. A., Drabick, D. A. G., Burgers, D. E., & Rabinowitz, J. (2015). The validity of the multi-informant approach to assessing child and adolescent mental health. *Psychological Bulletin, 141*(4), 858-900. http://dx.doi.org/10.1037/a0038498

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results, *5*, 1–17. https://doi.org/10.3389/fpsyg.2014.00781

de Vicente, M., Ballester Brage, L., Orte Socias, M. del C., & Amer Fernandez, J. A. (2017). Meta-analysis of family-based selective prevention programs for drug consumption in adolescence. *PSICOTHEMA*, *29*(3), 299–305. https://doi.org/10.7334/psicothema2016.275

Foxcroft, D. R., Ireland, D., Lowe, G., & Breen, R. (2003). Longer-term primary prevention for alcohol misuse in young people : a systematic review, 397–411.

Foxcroft, D.R., & Tsertsvadze, A. (2011). Universal school-based prevention programs for alcohol misuse in young people (Review), (5). https://doi.org/10.1002/14651858.CD009113.www.cochranelibrary.com

Frederickson, N. (2002). Evidence-based practice and educational psychology. *Educational and Child Psychology, 19*(3), 96-111.

Gates, S., Mccambridge, J., La, S., & Foxcroft, D. (2006). Interventions for prevention of drug use by young people delivered in non-school settings (Review). https://doi.org/10.1002/14651858.CD005030.pub2

Goodman, A. and Gregg, P. (2010) Poorer children's educational attainment: how important are attitudes and behaviour? London, Institute for Fiscal Studies. (n.d.). Retrieved from https://www.jrf.org.uk/sites/default/files/jrf/migrated/files/poorer-children-education-full.pdf

Gorman, D. M. (2017). The decline effect in evaluations of the impact of the Strengthening Families Program for Youth 10-14 (SFP 10-14) on adolescent substance use. *Children and Youth Services Review*, *81*(March), 29–39. https://doi.org/10.1016/j.childyouth.2017.07.009

Gupta S. K. (2011). Intention-to-treat concept: A review. *Perspectives in clinical research*, *2*(3), 109-12.

Hahn, S., Puffer, S., Torgerson, D. J., & Watson, J. (2005). Methodological bias in cluster randomised trials, *8*, 1–8. https://doi.org/10.1186/1471-2288-5-10

Harden, A., & Gough, D. (2012). Quality and Relevance Appraisal. In D. Gough, S. Oliver, & J. Thomas (Eds.). An Introduction to Systematic Reviews, 153–178, London: Sage

Hawkins, J. D., Catalano, R. F., & Miller, J. Y. (1992). Risk and protective factors for alcohol and other drug problems in adolescence and early adulthood: Implications for substance abuse prevention. *Psychological Bulletin, 112*(1), 64-105. http://dx.doi.org/10.1037/0033-2909.112.1.64

Israel, B.A., Parker, E.A., Rowe, Z., Salvatore, A., Minkler, M., Lopez, J., & Halstead, S. (2005). Community-based participatory research: Lessons learned from the Centers for Children's Environmental Health and Disease Prevention Research. Environmental Health Perspectives, 113, 1463–1147

Knepper, P. (2007). Criminology and Social Policy. London: Sage Publishing

Kratochwill, T. R. (2003). Evidence-Based Practice: Promoting Evidence-Based Interventions in School Psychology. *School Psychology Quarterly,* 18(4), 389- 408.

Kumfer, K. L., Trunnel, E. P., & Whiteside, H. O. (1990). The biopsychosocial model: applications to the addictions field. In: Engs RC editor. Controversies in the Addictions Field. Dubuque: Kendall Hunt Publishing.

Mejia, A., Ulph, F., & Calam, R. (2016). The Strengthening Families Program 10-14 in Panama: Parents' Perceptions of Cultural Fit. *PROFESSIONAL PSYCHOLOGY-RESEARCH AND PRACTICE*, *47*(1), 56–65. https://doi.org/10.1037/pro0000058

Molgaard, V. M., Spoth, R., & Redmond, C. (2000). Competency training: The Strengthening Families Program for Parents and Youth 10-14. OJJDP Juvenile Justice Bulletin (NCJ 182208). Washington, DC: U.S. Department of Justice, Office of Juvenile Justice and Delinquency Prevention

Molgaard, V., & Spoth, R. (2001). The Strengthening Families Program for young adolescents: Overview and outcomes. *Special Issue: Innovative Mental Health Interventions for Children: Programs That Work*, *18*(3), 15–29. https://doi.org/http://dx.doi.org/10.1300/J007v18n03_03

Molgaard, V., Spoth, R., & Molgaard, V. (2008). Residential Treatment for Children & Youth The Strengthening Families Program for Young Adolescents : Overview and Outcomes The Strengthening Families Program for Young Adolescents : Overview and Outcomes, *0358*. https://doi.org/10.1300/J007v18n03

Ortega, E., Giannotta, F., Latina, D., & Ciairano, S. (2012). Cultural Adaptation of the Strengthening Families Program 10-14 to Italian Families. *Child and Youth Care Forum*, *41*(2), 197–212. https://doi.org/10.1007/s10566-011-9170-6

Perez, J., Dıaz, S., Villa, R., Hermida., Crespo, J., & Rodriguez, G. (2012) Family-based drug use prevention: The "familias que funcionan" program. *Psychology Spain*, *14*(1), 1–7

Ross, A., Duckworth, K., Smith, D., Wyness, G., & Schoon, I. (2010). Prevention and Reduction: A review of strategies for intervening early to prevent or reduce youth crime and anti-social behaviour. Retrieved from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/182548/DFE-RR111.pdf

Rothbaum, F, & Weisz, J. R. (1994). Parental caregiving and child externalizing behaviour in non-clinical samples: a meta-analysis. *Psychology Bulletin, 116*(1), 55-74.

Roulette, J. W., & Hill, L. G. (2017). Cultural adaptations of the Strengthening Families Programme 10-14 in the US Pacific Northwest : A qualitative evaluation. https://doi.org/10.1177/0017896916654726

Skinner, E. A., & Kindermann, T. A. (2008). Academic Activities in the Classroom, (527594), 1–33. https://doi.org/10.1177/0013164408323233

Spoth, R. L., Redmond, C., & Shin, C. (2000). Reducing adolescents' aggressive and hostile behaviors - Randomized trial effects of a brief family intervention 4 years past baseline. *ARCHIVES OF PEDIATRICS & ADOLESCENT MEDICINE*, *154*(12), 1248–1257. https://doi.org/10.1001/archpedi.154.12.1248

Spoth, R. L., Randall, G. K., Trudeau, L., Shin, C., & Redmond, C. (2008). Substance use outcomes 5 1/2 years past baseline for partnership-based, family-school preventive interventions. *DRUG AND ALCOHOL DEPENDENCE*, *96*(1–2), 57–68. https://doi.org/10.1016/j.drugalcdep.2008.01.023

Spoth, R., Trudeau, L., Redmond, C., & Shin, C. (2016). Replicating and Extending a Model of Effects of Universal Preventive Intervention During Early Adolescence on Young Adult Substance Misuse. *Journal of consulting and clinical psychology*, *84*(10), 913–921. https://doi.org/10.1037/ccp0000131

Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in Schools, 44*(5), 423-432

APPENDIX A

Studies excluded from the review, including rationale.

Study	Rationale for Exclusion (criteria in brackets)
Jalling, C., Bodin, M., Romelsjo, A., Kallmen, H., Durbeej, N., & Tengstrom, A. (2016). Parent Programs for Reducing Adolescent's Antisocial Behavior and Substance Use: A Randomized Controlled Trial. *JOURNAL OF CHILD AND FAMILY STUDIES*, *25*(3), 811–826. https://doi.org/10.1007/s10826-015-0263-y	(4) Participants aged 12-18
Kumpfer, K. L., Xie, J., & O'Driscoll, R. (2012). Effectiveness of a Culturally Adapted Strengthening Families Program 12-16-Years for High-Risk Irish Families. *Child & Youth Care Forum*, *41*(2), 173–195. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=eric&AN=EJ959328&site=ehost-live&scope=site	(4) Participants aged 12-16 years
Spoth, R. L., Redmond, C., & Shin, C. (2000). Reducing adolescents' aggressive and hostile behaviors - Randomized trial effects of a brief family intervention 4 years past baseline. *ARCHIVES OF PEDIATRICS & ADOLESCENT MEDICINE*, *154*(12), 1248–1257.https://doi.org/10.1001/archpedi.154.12.1248	(5) Long term outcomes reviewed and reported (4 years past baseline). SFP 10-14 data subsumed within narrative about long term effects of SFP
Chartier, K. G., Negroni, L. K., & Hesselbrock, M. N. (2010). Strengthening Family Practices for Latino Families. Journal of Ethnic & Cultural Diversity in Social Work, 19(1), 1–17. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=eric&AN=EJ881419&site=ehost-live&scope=site	(3) Study design was qualitative
Coombes, L., Allen, D., Marsh, M., & Foxcroft, D. (2009). The Strengthening Families Programme (SFP) 10-14 and Substance Misuse in Barnsley: The Perspectives of Facilitators and Families. *CHILD ABUSE REVIEW*, *18*(1), 41–59. https://doi.org/10.1002/car.1055	(3) Study design was qualitative

Study	Rationale for Exclusion (criteria in brackets)
Broning, S., Sack, P.-M., Thomsen, M., & Thomasius, R. (2016). Children with multiple risk factor exposition benefit from the German “Strengthening Families Program”. *Kinder Mit Multipler Risikoexposition Profitieren von Der Teilnahme an Familien Starken"!*, *65*(7), 550–566. https://doi.org/http://dx.doi.org/10.13109/prkk.2016.65.7.550	(5) Study reports results derived from an alternative interpretation of a data set already reviewed in the main analysis
Lindsay, G., Strand, S., & Davis, H. (2011). A comparison of the effectiveness of three parenting programmes in improving parenting skills, parent mental-well being and children’s behaviour when implemented on a large scale community settings in 18 English local authorities: the parenting early intervention pathfinder (PEIP). *BMC Public*, 1–13. Retrieved from http://bmcpublichealth.biomedcentral.com/articles/10.1186/1471-2458-11-962	(2) Intervention is an alternative version of SFP 10-14
Bröning, S., Baldus, C., Thomsen, M., Sack, P. M., Arnaud, N., & Thomasius, R. (2017). Children with Elevated Psychosocial Risk Load Benefit Most from a Family-Based Preventive Intervention: Exploratory Differential Analyses from the German “Strengthening Families Program 10–14” Adaptation Trial. *Prevention Science*, *18*(8), 932–942. https://doi.org/10.1007/s11121-017-0797-x	(5) Study reports results derived from an alternative interpretation of a data set already reviewed in the main analysis
Ortega, E., Giannotta, F., Latina, D., & Ciairano, S. (2012). Cultural Adaptation of the Strengthening Families Program 10-14 to Italian Families. *Child and Youth Care Forum*, *41*(2), 197–212. https://doi.org/10.1007/s10566-011-9170-6	(3) Study design was qualitative
Kumpfer, K. L., Whiteside, H. O., Greene, J. A., & Allen, K. C. (2010). Effectiveness Outcomes of Four Age Versions of the Strengthening Families Program in Statewide Field Sites. *GROUP DYNAMICS-THEORY RESEARCH AND PRACTICE*, *14*(3, SI), 211–229. https://doi.org/10.1037/a0020602	(3) Study design explored long term effectiveness across different age versions of SFP as a repeated measures design

Appendix B

Full description of Studies Included for Review

Author	Intervention SFP 10-14 Replicated by:	Sample	Study Design	Participant (age, gender, presenting problem behaviour)	Measures at pre-test	Post test	Follow-Up
Baldus et al., 2016	Country: Germany (4 communities)	N=292 (intervention: n=147)	RCT (minimal interventio n)	10-14 years (M = 12.06); 121 girls and 171 boys; assessed as having problem behaviours N=3 children excluded for disruptive behaviours (seeking compatibility between group setting and its members	Child Self-report: Reynolds Adolescent Adjustment Screening (RAASI. Hempel et al., 2006) Scale 'Anti-social behaviour' >85th Percentile of norm. Validated for target population Parent Report: Strengths and Difficulties (SDQ, Goodman et al., 1998) >93rd percentile of norm	8 weeks	6 and 18 month s
Coombes et al., 2012	Country: UK (3 communities)	N=69 (intervention: n=34)	Quasi-experiment al (no interventio n)	10-14 years (M=11.2); 33 girls and 31 boys; no presenting problem behaviours 90% white british No group differences	Child Self-report: Questions from in-home national survey to assess frequency of aggressive behaviours Parent and Child self –report: Extraction from validated Iowa Youth and Family Rating Scales on Perceptions of Hostility/Warmth to assess adolescent aggressive and hostile behaviors in parent-adolescent interactions (Spoth, Redmond, & Shin, 2000)	Final Session	3 month s

Author	Intervention SFP 10-14 Replicated by:	Sample	Study Design	Participant (age, gender, presenting problem behaviour)	Measures at pre-test	Post test	Follow-Up
Foxcroft at al., 2016	Country: Poland 20 communities. Recruitment through schools, community agencies, information leaflets and personal contact	614 (intervention n=367)	RCT (minimal contact)	10-14 years (M=12.1); girls 358 and 256 boys	Child Self-report: Aggressive and Hostile Behaviours in Interactions; Aggressive and Destructive Conduct validated (Spoth et al., 1998); SDQ (externalising sub-scale).	12 months	24 months
Skarstrand et al., 2013	Country: Sweden All elementary schools in Stockholm invited	587 (intervention n=371)	RCT (waitlist)	10-14 (M=12); 292 girls and 295 boys	Child Self-Report 'Norm-Breaking Behaviours' Scale (15 item scale – cronbachs .86)	12 months	24 and 48 months
Spoth et al., 2003	Race: African American Drawing on census data, public schools in	110 (from initial 200; w/ 85 providing data for analysis;	RCT (waitlist)	10-14 years (M=10.5); girls and boys (N/A)	Child Self-report: Child Behaviour Scale validated from Spoth et al., 1998 (Cronbach's alpha .70)	4 weeks	8 weeks

	Iowa approached	intervention n= 34)

Appendix C

Weighting of Studies

WoE A: Methodological Quality

Establishing the methodological quality of a study requires a generic judgement of its research design. This is widely accepted as an evaluation against a criteria established through published protocols. For this review, the Kratochwill Group-Based Design Protocol was adapted and applied. The protocol is defined in two sections. The first allows appropriateness of fit to be determined, whilst the second section provides main dimensions of the methodological quality of a study. Dimensions are mapped against criteria in the Coding Manual, which are then used to inform judgements of these dimensions. Points are awarded and given qualitative descriptions as follows: No Evidence = 0; Weak Evidence = 1; Promising Evidence = 2; and Strong Evidence = 3. These scores are reported by dimension at the end of each protocol, and are summarised in the table below.

Included Studies: Rating per section of Kratochwill (2003) coding protocol

Main Dimensions	Baldus et al., (2016)	Coombes et al., (2012)	Foxcroft et al., (2017)	Skarstrand et al., (2014)	Spoth et al., (2003)
Measurement	3	3	1	1	2
Comparison Group	2	1	1	3	1
Fidelity	3	3	3	3	3
Replication	3	3	3	3	2
Follow Up	3	3	3	3	2

An overall evidence rating is reached by calculating a mean score. Upper and Lower markers are split into terciles and used to reflect the range; scored and rated as follows: Low = 0.0 to 1.0; Medium = 1.1 to 2.0; and High = 2.1-3.0. These scores and ratings are summarised below to reflect the Methodological Quality.

Weight of Evidence A – Score and Rating

Study	Mean Score	Quality Rating
Baldus et al., (2016)	2.8	High
Coombes et al., (2012)	2.6	High
Foxcroft et al., (2017)	2.2	High
Skarstrand et al., (2014)	2.6	High
Spoth et al., (2003)	2.0	Medium

WoE B: Methodological Relevance to the Review Question

The methodological relevance to the review question is evaluated by making review-specific judgements about the appropriateness of the study evidence to answer the review question. The following three elements were considered to be especially methodologically relevant:

1. Randomised controlled studies to establish efficacy
2. Suitable control group to establish the effectiveness of an adapted version of the age specific intervention: SFP:10-14
3. Measures used to report outcomes of problem behaviours in children aged 10-14 years
4. Implementation fidelity to the original intervention; with exception to adaptions otherwise approved by the original intervention designers.

Weighting criteria for appraising WoE B for included studies

Weighting	Description
High (3 points)	1. All participants should be randomly assigned to either the intervention or control group. Equivalence should be checked and reported 2. The control group is provided with the original intervention SFP:10-14 not adapted 3. Child problem behaviours which are recognised and published as likely 'risk factors' must be measured and reported (e.g. SDQ). These can be internalised or externalised behaviours

	4. The intervention must follow the original intervention format except for adaptions otherwise approved
Medium (2 points)	1. Participants are non-randomised to either the intervention or control group with group equivalence established and reported 2. The control group is provided with wait-list or delayed intervention; minimal contact; or discussion 3. Other child problem behaviours are measured and reported which are not necessarily recognised as a mediating 'risk factor' (e.g. cheating on a test or other miscellaneous 'norm-breaking' behaviours) 4. The intervention mostly follows the original format as well as including otherwise approved adaptions
Low (1 point)	1. Non-randomised design with no group equivalence not calculated corrected for, or reported. 2. The control group is given an alternative intervention which is widely considered to have an effect (but is not SFP:10-14) 3. Other non-problem child behaviours are measured and reported 4. The intervention does not follow the original format even as approved adaptions are implemented

An overall evidence rating is reached by calculating a mean score. Upper and Lower markers are split into terciles and used to reflect the range; scored and rated as follows: Low = 1-1.6 Medium = 1.7to 2.3; and High = 2.4-3.0. These scores and ratings are summarised below to reflect the Methodological Relevance to the review question.

Scores for WoE B for each included study across each criteria separately, and including mean score (in brackets) and overall Quality rating

Study	Score for each criteria	Mean Score Quality Rating
Baldus et al., (2016)	1 = High (3) 2 = Medium (2) 3 = High (3) 4 = High (3)	(2.8) High
Coombes et al., (2012)	1 = Medium (2) 2 = Medium (2) 3 = High (3) 4 = Medium (2)	(2.25) Medium

Foxcroft et al., (2017)	1 = High (3) 2 = Medium (2) 3 = High (3) 4 = High (3)	(2.8) High
Skarstrand et al., (2014)	1 = High (3) 2 = Low(1) 3 = Medium (2) 4 = Low (1)	(1.75) Medium
Spoth et al., (2003)	1 = High (3) 2 = Medium (2) 3 = High (3) 4 = Medium (2)	(2.5) Medium

WoE C. Topic Relevance to the Review Question

To determine whether the focus and character of the study contributed towards answering the review question, the following elements were appraised as having topic relevance:

1. Children should be recruited according to existing presentation of problem behaviour difficulties such that the effectiveness on reducing child problem behaviours can be demonstrated.

2. Rationale and discussion of the study is specific to the adaption of the intervention and effectiveness on child problem behaviours

3. Parent and child self-report are both included. This will likely mitigate for contention in the literature as to the most valid measure of child problem behaviours

Weighting criteria for WoE C for appraising included studies

Weighting	Description
High (3 points)	1. Children are recruited to the study with pre-existing problem behaviour difficulties or are appraised as 'at risk' at the start of the study 2. The adaption of the intervention clearly informs the rationale for the study, and is expressly contextualised within the discussion, including interpretation of results 3. Parent and child self-report are both included

Medium (2 points)	1. Only the parent or local community is assessed as being 'at risk' at the start of the study 2. The adaption of the intervention partly informs the rationale and discussion, including interpretation of results 3. Parent or child self-report are included
Low (1 point)	1. Neither children, parents, or the local community are assessed as being 'at risk' at the start of the study 2. The adaption of the intervention does not inform the rationale nor discussion, including interpretation of results. 3. Neither parent or child self-report are included

An overall evidence rating is reached by calculating a mean score of the three elements. Upper and Lower markers are split into terciles and used to reflect the range; scored and rated as follows: Low = 1-1.6 Medium = 1.7 to 2.3; and High = 2.4-3.0. These scores and ratings are summarised below to reflect the Topic Relevance to the help answer the review question.

Scores for WoE C for each included study across each criteria separately, and including mean score (in brackets) and overall Quality rating

Study	Score for each criteria	Mean Score Quality Rating
Baldus et al., (2016)	1. Medium = 2 2. High = 3 3. High = 3	(2.7) High
Coombes et al., (2012)	1. Low = 1 2. High = 3 3. Medium = 2	(2) Medium
Foxcroft et al., (2017)	1. Low = 1 2. Medium = 2 3. Medium = 2	(1.7) Medium
Skarstrand et al., (2014)	1. Low = 1 2. High = 3 3. Medium = 2	(2) Medium
Spoth et al., (2003)	1. Medium = 2 2. High = 3 3. High = 3	(2.7) High

Appendix D

Rationale for Adaptation of Kratochwill's (2003) Coding Protocol

Note: For author ease of use, this review applied its own alphabetised and numerical order to Kratochwill's (2003) coding protocol. Please see Appendix E for the individual coding protocols of included studies.

Adaptions made to Kratochwill (2003) coding protocol including items removed and rationale for doing so.

Sections of protocol removed	Rationale
Part I B. Statistical Treatment/Data Analysis B7 - B8	Evaluating the effectiveness of SFP:10-14 required quantitative data, this section relates to qualitative coding and was considered not relevant
Part II. Key Features for Coding Studies and Rating Level of Evidence/Support C2. C2.1-C2.9 Primary Outcomes Are Statistically Significant C3-C5 Secondary Outcomes	This review only evaluation secondary outcomes. To avoid duplication, this was reviewed separately.
Part II. Key Features for Coding Studies and Rating Level of Evidence/Support D. Educational/Clinical Significance	The type of intervention evaluated in this review meant this section was not relevant. The intervention was not specific to educational settings, and participants were not considered to meet any clinical threshold such that outcomes would have clinical relevance.
Part II. Key Features for Coding Studies and Rating Level of Evidence/Support E. Identifiable Components	Whilst valuable for future research, this section was not considered relevant to this review. Some elements are evaluated and summarised separately.

Part III. Other Descriptive or Supplemental Criteria to Consider A. External Validity Indicators A4. Receptivity/acceptance by target participant population (treatment group)	Included separately in tables summarising the intervention more fully. Removed from coding to avoid duplication.
Part III F. Characteristics of Intervener G. Intervention style or orientation H. Cost Analysis Data I Training and Support Resources J Feasibility .	Insufficient information available. Some elements have been discussed separately as appropriate.

Appendix E

Kratochwill Coding Protocol Example

[Adapted from the Procedural Manual of the Task Force on Evidence-Based Interventions in School Psychology, American Psychology Association, Kratochwill, T.R. (2003)]

Coding Protocol

Name of Coder: Anon **Date:** 13.01.2019

Full Study Reference in proper format:

Baldus, C., Thomsen, M., Sack, P. M., Bröning, S., Arnaud, N., Daubmann, A., & Thomasius, R. (2016). Evaluation of a German version of the Strengthening Families Programme 10-14: A randomised controlled trial. *European Journal of Public Health*, *26*(6), 953–959. https://doi.org/10.1093/eurpub/ckw082

Intervention name: (description of study):
Strengthening Families Programme 10-14 (German version)

Study ID number: 1.

Type of Publication:

☐ Book/Monograph
☒ Journal Article
☐ Book Chapter
☐ Other (specify):

I. General Characteristics

A. General Design Characteristics

A1. Random assignment designs (if random assignment design, select one of the following)

☐ Completely randomized design
☒ Randomized block design (between participants, e.g., matched classrooms)
☐ Randomized block design (within participants)
☐ Randomized hierarchical design (nested treatments

A2. Nonrandomized designs (if non-random assignment design, select one of the following)

☐ Nonrandomized design
☐ Nonrandomized block design (between participants)
☐ Nonrandomized block design (within participants)
☐ Nonrandomized hierarchical design
☐ Optional coding for Quasi-experimental designs

A3. Overall confidence of judgment on how participants were assigned (select one of the following)

☐ Very low (little basis)
☐ Low (guess)
☐ Moderate (weak inference)
☐ High (strong inference)
☒ Very high (explicitly stated)
☐ N/A
☐ Unknown/unable to code

B. Statistical Treatment/Data Analysis (answer B1 through B6)

	Yes	No
B1. Appropriate unit of analysis	☒	☐
B2. Familywise error rate controlled	☒	☐
B3. Sufficiently large N	☒	☐

Total size of sample (start of study): n=292 *(power calculated based on previous medium size effects from original studies(n=288 at baseline).*

Intervention group sample size n= 153

Control group sample size: n= 149

C. Type of Program

- ☐ Universal prevention program
- ☒ Selective prevention program
- ☐ Targeted prevention program
- ☐ Intervention/Treatment
- ☐ Unknown

D. Stage of Program (select one)

- ☒ Model/demonstration programs
- ☐ Early stage programs
- ☐ Established/institutionalized programs
- ☐ Unknown

E. Concurrent or Historical Intervention Exposure (select one)

- ☐ Current exposure
- ☐ Prior exposure
- ☒ Unknown

I. Key Features for Coding Studies and Rating Level of Evidence/Support

(Rating Scale: **3= Strong Evidence, 2=Promising Evidence, 1=Weak Evidence, 0=No Evidence)**

A. Measurement (answer A1 through A4)

A1. Use of outcome measures that produce reliable scores for the majority of (secondary) outcomes.

- ☒ Yes
- ☐ No
- ☐ Unknown/unable to code

A2 Multi-method (select one of the following)

- ☐ Yes
- ☒ No
- ☐ N/A
- ☐ Unknown/unable to code

A3 Multi-source (select one of the following.)

- [x] Yes
- [ ] No
- [ ] N/A
- [ ] Unknown/unable to code

A4 Validity of measures reported (select one of the following)

- [x] Yes validated with specific target group
- [ ] In part, validated for general population only
- [ ] No
- [ ] Unknown/unable to code

Rating for measurement (select 0, 1, 2 or 3) [x] 3 [ ] 2 [ ] 1 [ ] 0

B. Comparison Group

B1 Type of Comparison Group (Select one of the following)

- [ ] Typical contact
- [ ] Attention placebo
- [ ] Intervention element placebo
- [x] Alternative intervention
- [ ] Pharmacotherapy
- [ ] No intervention
- [ ] Wait list/delayed intervention
- [ ] Minimal contact
- [ ] Unable to identify type of comparison

B2 Overall confidence of judgment on type of comparison group

- [ ] Very low (little basis)
- [ ] Low (guess)
- [ ] High (strong inference)
- [x] Very high (explicitly stated)
- [ ] Unable to identify comparison group

B3 Counterbalancing of change agent

- [ ] By change agent
- [x] Statistical (analyse includes a test for intervention) Supplementary Data provided and reviewed
- [ ] Other
- [ ] Not reported/None

B4 Group equivalence established (select one of the following)

- [x] Random assignment
- [ ] Posthoc matched set
- [ ] Statistical matching
- [ ] Post hoc test for group equivalence

B5 Equivalent mortality

☐ Low attrition (less than 20 % for post)
☐ Low attrition (less than 30% for follow-up)
☐ Intent to intervene analysis carried out?
Findings_____

Overall rating for Comparison group (select 0, 1, 2 or 3) ☐ 3 ☐2 ☐ 1 ☐ 0

C. Implementation Fidelity

C1. Evidence of Acceptable Adherence

☐ Ongoing supervision/consultation
☐ Coding intervention sessions/lessons or procedures
☐ Audio/video tape implementation
☐ Entire intervention
☐ Part of intervention. *Supplementary Data provided and was reviewed Adherence % is M = 85.5 (CI 95%; 83.6 – 88.0), that is 85.5% of the manual's contents was delivered; no inter-rater agreement was computed. Adherence was not statistically significant different between locations (4 sites), session type (children–parents–family), time clips (initial–mid-session–final), and session number (session no. 4 or 5).*

F2. Manualization (select all that apply)

☐ Written material involving a detailed account of the exact procedure and the sequence they are to be used.
☐ Formal training session that includes a detailed account of the exact procedures and the sequence in which they are to be used.
☐ Written material involving an overview of broad principles and a description of the intervention phases.
☐ Formal or informal training session involving an overview of broad principles and a description of the intervention phases.

Rating for Implementation Fidelity (select 0, 1, 2 or 3): ☐ 3 ☐ 2 ☐ 1 ☐ 0

D. Replication (answer D1, D2,

☐ Same Intervention
☐ Same Target Problem
☐ Independent evaluation

Rating for Replication (select 0, 1, 2, or 3)**:** ☐ **3** ☐ **2** ☐ **1** ☐ **0**

E. Follow-Up Assessment

☐ Timing of follow up assessment: <u>6 months and 18 months</u>

☐ Number of participants included in the follow up assessment: 6 months: Intervention n=136 and Control n=132. 18months: Intervention n=132 and control n=127

☐ Consistency of assessment method used: Same measures administered

Rating for Follow Up Assessment (select 0, 1, 2, or 3): ☐ **3** ☐ **2** ☐ **1** ☐ **0**

F. Additional Criteria and Supplementary information for Consideration

F.I Session duration

- [x] F1. Approved 7 weeks
- [ ] F2. Other number of sessions
- [ ] F3. Not specified

F.II. Program Implementer (Approved F1 to F5)

- [ ] F1. Research Staff
- [x] F2. School or Specialty Staff
- [ ] F3. Teachers
- [ ] F4. Educational Assistants
- [ ] F5. Parents
- [ ] F6. College Students
- [ ] F7. Peers
- [ ] F8. Other (Volunteers)
- [ ] F9. Unknown/insufficient information provided

F.III Location of implementation (Approved Schools, Churches, Community Centre, Youth Centre or Social Services)

- [x] F1. Approved implementation location
- [ ] F2. Not approved location
- [ ] F3. Not specified.

Summary of Evidence

Indicator	**Overall evidence rating 0-3**	**Description of evidence Strong Promising Weak No/limited evidence** **Or Descriptive ratings**
General Characteristics		
Design		
Type of programme		
Stage of programme		
Concurrent/ historical intervention exposure		
Key features		

Measurement	3	Strong
Comparison group	2	Promising
Implementation Fidelity	3	Strong
Replicability	3	Strong
Follow-up Assessment	3	Strong