Case Study 1: An Evidence-Based Practice Review Report

Theme: School/Setting Based Interventions for Social, Emotional and Mental Health

How effective is SPARX (a CBT computer game) in reducing depressive symptoms for youth in educational settings?

Summary

SPARX is a novel CBT computer game developed as a universal and targeted intervention for adolescents to reduce depressive symptoms. This systematic literature review aims to evaluate how effective SPARX is in reducing depressive symptoms for youth in educational settings. A systematic literature review was undertaken using three online data bases. Five studies were selected for review and evaluated using Harden and Gough's (2012) Weight of Evidence Framework and the adapted APA Task Force Coding Protocol (Kratochwill, 2003). Four studies were randomised controlled trials and one was a within-subjects pre-post design. Effect sizes were calculated for impact of SPARX on depressive symptoms. Data were extracted on demographics, age of participants, country and context of intervention. One high quality study showed evidence of a small post-intervention effect of SPARX compared with comparison group on depressive symptoms which was not maintained at follow up. Two medium quality studies showed evidence of a large intervention effect of SPARX compared with comparison group or within-subjects, which was maintained at follow-up. Overall, the data provide some preliminary support in

favour of SPARX as an effective intervention for youth with depressive symptoms when delivered within educational settings. However, several methodological limitations are noted amongst the studies reporting the largest effects. Further higher quality randomised trials, with larger sample sizes and more diverse settings, are needed to inform whether widespread implementation of SPARX in educational settings is warranted.

Introduction

SPARX: A novel CBT computer game

SPARX (Smart, Positive, Active, Realistic, X-factor throughs) is an interactive fantasy-based computer game designed to enable adolescents to learn and practice cognitive behavioural therapy (CBT) principles to reduce depressive symptoms (Merry et al., 2012). At present, SPARX is one of very few available evidence-based gamified CBT interventions specifically developed for adolescents with depressive symptoms that does not require a facilitator or clinician to be completed (Lucassen et al., 2015).

SPARX was originally developed in New Zealand by clinicians and academics in partnership with a computer games company, with advice from young people and Maori, Pacific an Asian cultural advisors (Merry et al., 2012). Having first been developed as a treatment intervention for young people aged 12 to 19 years with mild-to-moderate symptoms of depression, it has also been adapted as a universal preventative intervention for adolescents (known as ‘SPARX-R’) and as a targeted intervention for Lesbian, Gay, Bisexual and Transgender (LGBT) youth (known as ‘Rainbow SPARX’) who are at increased risk of depressive symptoms (Almeida et al., 2009).

All three versions of SPARX are largely similar in terms of content, but each features a tailored script which makes the story more relevant as a targeted treatment approach for depression (SPARX), universal preventative approach for young people feeling down, angry or stressed (SPARX-R), or targeted for LGBT youth and the unique difficulties they face (Rainbow SPARX). It has recently been adapted and translated into Japanese (Yokomitsu et al., 2020) and is currently being redesigned for Nunavut youth (First Nations young people in Arctic Canada) (Bohr & Merry, 2016).

Format of Delivery and Main Features

SPARX is delivered in the format of seven sequential modules or levels, each of approximately 20-30 minutes duration, which can be accessed online or via a CDROM (Kuosmanen et al., 2017). The player controls an avatar across the seven distinct levels while interacting with characters who explain CBT concepts, teach the player skills and strategies, and give the player 'homework' to try these strategies while away from the game. Table 1 provides a summary of content within each module.

Table 1

Overview of SPARX modules, adapted from Lucassen et al., (2015)

Module	Main Content Covered
1 Cave Province: 'Finding Hope'	• Introducing unhelpful thoughts • The character and concept of Hope • Controlled breathing • Psycho-education about depression and the CBT model
2 Ice Province: 'Being Active'	• Progressive muscle relaxation • Communication skills • Behavioural activation and activity scheduling

3 Volcano Province: 'Dealing with Emotions'	• Listening skills • Identifying strong emotions
4 Mountain Province: 'Overcoming Problems'	• Introducing problem solving • Recognising sparks (positive or helpful thoughts about you/your future)
5 Swamp Province: 'Recognising Unhelpful Thoughts'	• Recognising various negative automatic thoughts
6 Bridgeland Province: 'Challenging Unhelpful Thoughts	• Learning to challenge negative automatic thoughts
7 Canyon Province: 'Bringing it all together'	• Recap of all skills • Mindfulness • Knowing when to ask for help

Mechanism of Change

SPARX is considered a 'serious game', in which education and behaviour change is the goal, alongside entertainment (Cheek et al., 2015). Serious games based in theory, evidence, and tailored to psychological constructs have been found to contribute to increased adherence and efficacy.

SPARX was developed using cognitive behavioural therapy (CBT) and learning theory, with input on game design from youth and stakeholders (Cheek et al., 2015). CBT is a structured, short-term psychological therapy (Beck, 1995), and evidence suggests that it is an effective intervention for adolescent depression (Wanatabe et al., 2007).

The mechanism of change behind CBT for depression is based on Beck's (1976) cognitive theory which suggests that the way we think (cognition) about situations can affect the way we feel (emotion) and how we act (behaviour). According to Beck (1976), mood change is not due to the situation itself, but the

negative automatic thoughts that the situation elicited and the cognitive distortions which maintain a depressed mood. Therefore in CBT, individuals learn to address maladaptive behaviours and psychological distress by altering the cognitive processes and behaviours that sustain them (Beck, Rush, Shaw and Emery, 1979).

In each module of the SPARX intervention, users are explicitly introduced to core components of CBT for depression including psychoeducation, relaxation skills, activity scheduling, problem solving, cognitive restructuring, interpersonal skills, help seeking, and dealing with strong emotions (Merry et al., 2012).

These skills are initially taught using a virtual therapist or guide, after which the user transitions to a fantasy setting to undertake CBT-based challenges and develop CBT-based skills within an overall narrative of restoring balance to the fantasy world. Following this exploratory learning, users return to the guide at the end of each level to reflect on the tasks and how they might be applied in their own lives.

Self-determination theory (SDT) proposes that when individuals perceive they have more control over their treatment, a sense of competence in the activities and tasks required of them, and a sense of being cared for and connected with another, they are more likely to integrate learning and behaviour change (Rogers, 1975). SDT is likely to be a useful theory in understanding the mechanisms of change for youth engaging with SPARX.

Benefits of SPARX

The efficacy of SPARX as a clinical treatment for young people seeking help for low mood or depression has been demonstrated in a large randomised controlled non inferiority trial (RCT) (Merry et al., 2012), which reported pre- to post-intervention decreases in depressive symptoms for adolescents using SPARX, and equivalent outcomes between SPARX and face-to-face therapy (usual care).

Various qualitative studies of youth trialling or engaging with SPARX have reported high levels of satisfaction with the programme. Themes included valuing the choices and control SPARX offered, the accessibility of the game and how it protected their privacy. Increased engagement with the programme has been linked to the playful medium, the ability for users to customise their own character, having a perceived sense of benefit from the programme, and interacting with characters that instil the user with a sense of care and hope (Cheek et al., 2014; Fleming et al., 2016; Fleming et al., 2012; Lucassen et al., 2013; Shepherd et al., 2015).

Rationale for Review & Relevance to Educational Psychology Practice

Integrating a range of accessible user-driven options into general community-level settings is one of the strategies promoted in the World Health Organization Mental Health Action Plan. There is an emphasis on early intervention and autonomy, particularly for young people (WHO, 2020).

Computerised CBT (CCBT) has been shown to significantly reduce symptoms of anxiety and depression in youth (Calear & Christensen, 2010) and is promising due to low cost and appeal to adolescents used to modern technology (Poppelaars et al., 2016).

There is no current systematic synthesis of the effectiveness of SPARX in educational settings. This review therefore seeks to understand whether SPARX holds promise as an effective intervention for reducing depressive symptoms for youth when delivered through the educational setting context, such as during class time and with minimal supervision.

SPARX may have educational significance given the flexibility of its implementation and ability to extend the reach of standardised CBT to more young people, including those in the community whose access to support is limited by barriers inherent to depression itself, such as reduced motivation and help-seeking behaviour.

Digitalised SEMH interventions are particularly relevant for EP practice at the current time in which face-to-face practice is limited due to health and safety restrictions during COVID-19. The socio-emotional and mental health needs of young people continue to rise well beyond the capacity of current service provision, and questions remain as to how this can be addressed within education in the context of the pandemic and beyond (DfE, 2020; The Children's Society, 2020).

Review question

The primary aim of the current review is to address the question:

How effective is SPARX in reducing depressive symptoms for youth within educational settings?

Critical Review of the Evidence Base

Literature search

A literature search was conducted on the 14th of January 2021 via PsycInfo (psychology database), ERIC (Proquest) (educational database) and Web of Science (interdisciplinary database), using the search terms listed in Table 2 to identify potentially relevant studies. Medline was also searched given the topic focus of depression and its relevance to medical literature, however only duplicates were found.

Table 2

Search Terms

Databases searched	**Search Terms**
PsychINFO	SPARX*
ERIC (Proquest)	OR
Web of Science	'Rainbow SPARX'

The resulting 64 studies were initially screened by title and abstract. If further clarification was needed the full text was reviewed. The criteria used to screen the studies are presented in Table 3, which details the inclusion and exclusion criteria used for this review.

Table 3

Inclusion and Exclusion Criteria

		Inclusion criteria	**Exclusion criteria**	**Rationale**
1	Participants	Participants are aged between 10-24 years old	Participants younger than 10 or older than 24	SPARX was developed for adolescents and youth. 'Adolescence' is broadly defined as beginning and ending with pubertal transitions (Blakemore

				et al., 2010). 'Youth' is broadly defined as the 15-24 year age group (WHO, 2021)
2	Setting	SPARX must be delivered in an educational setting	SPARX not delivered in an educational setting	SPARX has not been reviewed in educational settings before
3	Type of Intervention	Study must have delivered a variant of SPARX intervention	Study did not deliver a variant of SPARX intervention	To allow the reviewer to critically evaluate the effectiveness of SPARX
4	Outcomes	One of the primary outcome measures are depressive symptoms	None of the primary outcome measure are depressive symptoms	To review the effectiveness of SPARX on depressive symptoms
5	Research design and methodology	Empirical data derived from quantitative studies with pre- and post-measures	Empirical data derived from qualitative studies or those missing quantitative data	Empirical data is sought for effect size calculation. An experimental design with pre- and post-measures is required to determine intervention impact
6	Type of publication	Studies published in an accessible peer-reviewed journal	Studies not published in an accessible peer-reviewed journal	Accessibility and peer-review is considered a minimum threshold of study quality and inclusion for this review

Figure 1

Study Selection Flow Chart

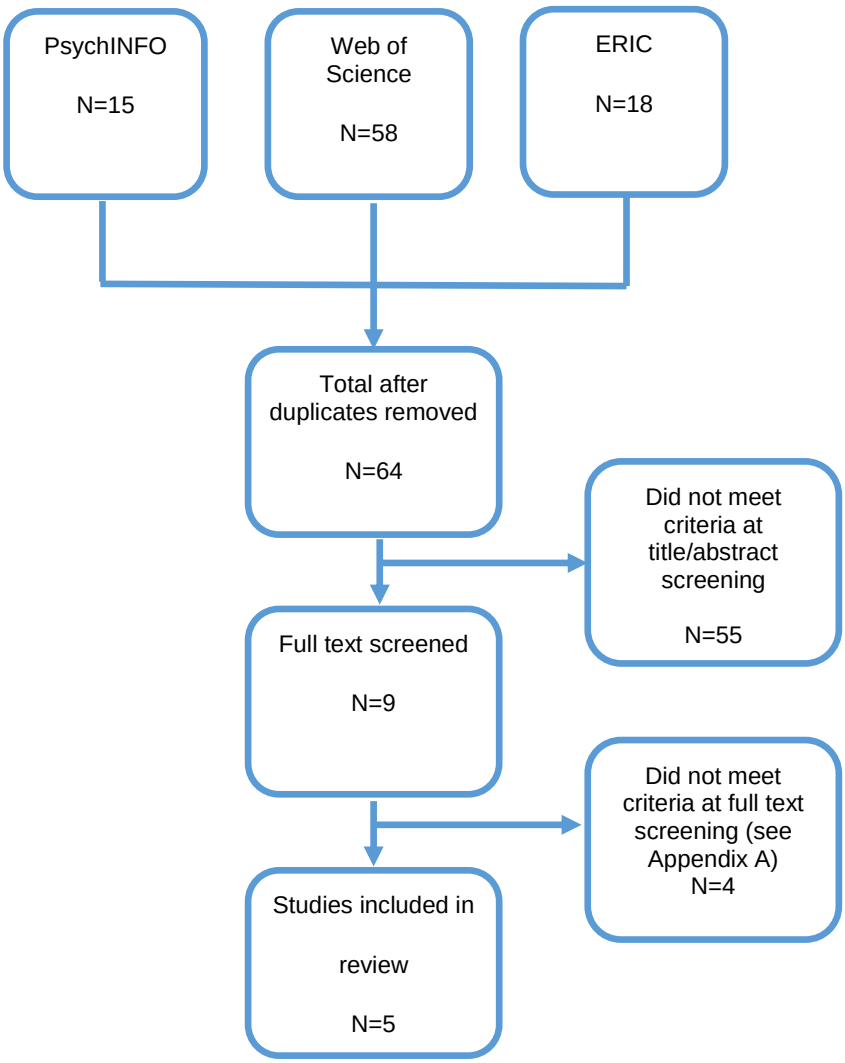

Table 4

Studies included in the review

Studies
Fleming, T., Dixon, R., Frampton, C., & Merry, S. (2012). A Pragmatic Randomized Controlled Trial of Computerized CBT (SPARX) for Symptoms of Depression among Adolescents Excluded from Mainstream Education. *Behavioural and Cognitive Psychotherapy*, *40*(5), 529–541.
Kuosmanen, T., Fleming, T. M., Newell, J., & Barry, M. M. (2017). A pilot evaluation of the SPARX-R gaming intervention for preventing depression and improving wellbeing among adolescents in alternative education. *Internet Interventions*, *8*, 40–47.
Lucassen, M. F. G., Merry, S. N., Hatcher, S., & Frampton, C. M. A. (2015). Rainbow SPARX: A novel approach to addressing depression in sexual minority youth. *Cognitive and Behavioral Practice*, *22*(2), 203–216.
Perry, Y., Werner-Seidler, A., Calear, A., Mackinnon, A., King, C., Clin, M., Scott, J., Merry, S., Fleming, T., Stasiak, K., Christensen, H., & Batterham, P. J. (2017). Preventing Depression in Final Year Secondary Students: School-Based Randomized Controlled Trial. *Journal of Medical Internet Research*, *19*(11).
Poppelaars, M., Tak, Y. R., Lichtwarck-Aschoff, A., Engels, R. C., Lobel, A., Merry, S. N., Lucassen, M. F., & Granic, I. (2016). A randomized controlled trial comparing two cognitive-behavioural programs for adolescent girls with subclinical depression: A school-based program (Op Volle Kracht) and a computerized program (SPARX). *Behaviour Research and Therapy*, *80*, 33.

Weight of Evidence (WoE)

The selected studies were evaluated according to Gough's Weight of Evidence (WoE) framework (Harden & Gough, 2012). Each study was judged on methodological quality (WoE A), methodological relevance (WoE B), and topic relevance (WoE C). The scores for each of the three categories were summed and divided by three to give an overall Weight of Evidence rating for the study (WoE D).

Table 5 summarises the WoE scores given to each study in this review. An adapted version of Kratochwill et al.'s (2003) protocol was used to assess WoE A for group designs. The sections omitted or modified and the accompanying rationale are detailed in Appendix B. Full details of the weight of evidence ratings and coding protocols are contained within the appendices.

Table 5

Overview of WoE ratings

Study	WoE A Methodological Quality	WoE B Methodological Relevance	WoE C Topic Relevance	WoE D Overall Weight of Evidence (Mean of A,B & C)
Fleming et al., 2012	Medium 2.25	High 2.5	Medium 2	Medium 2.25
Kuosmanen et al., 2017	Medium 1.75	High 2.5	Medium 2	Medium 2.08
Lucassen et al., 2015	Medium 1.67	Medium 1.5	High 2.5	Medium 1.89
Perry et al., 2017	High 3	High 2.5	High 2.5	High 2.67
Poppelaars et al., 2016	High 3	High 3	High 3	High 3

Mapping the Field

Key information pertaining the five included studies' characteristics, participants, intervention and measures are provided in Table 6.

Table 6 *Mapping the Field*

Study	Study type	Location	Age of participants (years)	Participant characteristics	Intervention	Context of intervention	Outcome variables measured
Fleming et al., 2012	Pragmatic RCT (Wait list control) with 5 week post-intervention f/up	New Zealand	13-16	Male 56% 38% Pacific, 34% Maori, 25% New Zealand European, 1 'other' Students who had been excluded from mainstream schooling Majority of students with elevated depressive symptoms at baseline	SPARX (universal)	Alternative education sites Delivered with a minimum threshold of monitoring during scheduled class time; one module per week.	Child Depression Rating Scale Revised (Poznanski & Mokros, 1996), Reynolds Adolescent Depression Scale (Reynolds, 2002)
Kuosmanen et al., 2017	Pilot Cluster RCT (No intervention control)	Ireland	15-20	Male 48.9% Early school leavers attending an alternative education programme for second level qualifications and vocational training	SPARX-R (universal)	Alternative education sites Delivered with a minimum threshold of monitoring during scheduled class time; one module per week.	Short Moods and Feelings Questionnaire (Angold et al., 1995), Generalized Anxiety Disorder Rating Scale (Spitzer et al., 2006),Warwick-Edinburgh Mental Wellbeing Scale (Tennant et al., 2007), Coping Strategy Indicator (Ellis, 2004), Emotional Regulation Questionnaire (Gross & John, 2003)
Lucassen et al., 2015	Pilot Feasibility Trial (uncontrolled) with 3 m f/up	Auckland, New Zealand	13-19	Male 52.4% 71.4% New Zealand European, 14.3% Asian, 9.5% Maori, 4.8% Pacific Sexual minority youth Elevated depressive symptoms	Rainbow SPARX (targeted)	Educational and community settings (including home, a youth-led organisation for sexual minority youth and high schools) Delivered flexibly without monitoring in time/place of their choosing.	Child Depression Rating Scale Revised (Poznanski & Mokros, 1996), Reynolds Adolescent Depression Scale (Reynolds, 2002), The Mood and Feelings Questionnaire (Kent et al., 1997), Pediatric Quality of Life Enjoyment and Satisfaction Questionnaire (Endicott et al., 2006), Spence Children's Anxiety Scale (Spence, 1998), Kazdin Hopelessness Scale for Children (Kazdin et al., 1983)
Perry et al., 2017	Cluster RCT (Attention control) with 6 & 18 m f/up	Sydney, Australia	16-17	Male 48.3% Not Australian-born 80.6%, Home language not English 54.5% Final year secondary students	SPARX-R (universal)	Selective (98.3%) and non-selective (1.7%) government secondary schools prior to final exams. Delivered with minimum threshold of monitoring during scheduled class time	The Major Depression Inventory (Bech et al., 2001), Spence Children's Anxiety Scale (Spence, 1998), Youth Risk behaviour Survey (Brener et al., 1999), Depression Stigma Scale (Griffiths et al., 2008)
Poppelaars et al., 2016	RCT (Active Monitoring control) with 3,6 & 12 m f/up	Netherlands	11-16	Female 100% Netherlands-born 94.7% Elevated depressive symptoms at baseline	SPARX (targeted)	Secondary schools. Delivered flexibly without supervision in time/place of their choosing. Students provided with CDROM and could play game at home; asked to complete one level per week.	Reynolds Adolescent Depression Scale (Reynolds, 2002), Children's Depression Inventory (Craighead et al., 1998)

Participants

A total of 841 young people from Ireland, The Netherlands, Australia and New Zealand and were included in the current review, ages ranging from 11 to 20 years old (see Table 6).

Participants were recruited from a variety of educational settings including mainstream secondary schools (selective and non-selective) and alternative education schools/programmes related to exclusion (for students at-risk of exclusion, currently excluded, transitioning out of exclusion and who have left school early). In addition to educational settings, one sample of participants (Lucassen et al., 2015) were also recruited by a sexual minority youth-led organization.

All studies reported the gender ratio of students in their sample, which was relatively equal, with one female-only study sample (Poppelaars et al., 2016). All participants in one study were reported as identifying as sexual minority youth (adolescents attracted to same sex, both sex, or who are questioning their sexuality) (Lucassen et al., 2015).

However, further reporting of demographic information was limited or inconsistent across all studies. Of the reported information, high cultural diversity was indicated in two studies according to ethnicity and/or home languages spoken (Perry et al., 2017; Fleming et al., 2012), whereas low cultural diversity was indicated in two studies according to nationality and/or ethnicity (Poppelaars et al., 2016; Lucassen et al., 2015). High social or economic disadvantage was indicated by two studies according to known risk-factors associated with exclusion (Kuosmanen et al., 2017; Fleming et al., 2012).

Baseline depressive symptoms were elevated for all students in studies in which SPARX was used as a targeted intervention (Poppelaars et al., 2016; Lucassen et al., 2015) and the majority of students in one study in which SPARX was used as a universal intervention (Fleming et al., 2012). Depressive symptoms were also reported as 'at-risk' level in one further study in which SPARX was used as a universal intervention (Kuosmanen et al., 2017). Only one study reported baseline depressive symptoms that were below threshold for 'mild depression' (Perry et al., 2017) in which SPARX was delivered universally prior to final secondary exams.

Study Design

Four studies included in the current review were RCTs. Definitive RCTs were most favoured in WoE B due to the more robust causality inferences they contribute when determining effectiveness of SPARX. This resulted in the highest WoE ratings for both definitive RCTs (Perry et al., 2017; Poppelaars et al., 2016), followed by the two pilot or feasibility RCTs (Kuosmanen et al., 2017; Fleming et al., 2012). Finally, the non-randomised pre-post uncontrolled trial (Lucassen et al., 2015), received lowest WoE B rating due to limitations associated with uncontrolled trials (i.e., a lack of evidence of causality).

However, this review also aimed to determine the effectiveness of SPARX across diverse student populations and educational settings. Therefore, overall WoE B rating of the studies also considered generalisability of study findings, with multi-site studies and wide geographic coverage being most favoured. Poppelaars et al. (2016) received the highest overall WoE B rating due to it being a definitive RCT that was conducted across multiple schools spanning an entire country (Netherlands),

whereas Perry et al.'s (2017) definitive RCT was conducted in multiple schools but within a single city (Sydney).

Both of the pilot/feasibility RCTs scored favourably when generalisability was considered due to their multi-site, wide geographic coverage (across Ireland and New Zealand, respectively), while the non-randomised uncontrolled trial was conducted within a single city (Aukland),and therefore scored lowest on WoE B (with an overall weighted score bordering on the low range).

Methodological Quality

With respect to WoE A, the two definitive RCTs received the highest ratings across four dimensions (1) reliability of measures (≥.85 Cronbach's alpha), (2) comparison groups (including active attention and monitoring control groups; group equivalence by random assignment; Intention-to-treat (ITT) analyses (3) sufficient power (power calculated; sufficiently large N) and (4) follow-up assessments (multiple time points ranging from 3 to 18 months post intervention inviting all original participants).

In contrast, comparison groups lacked quality in both pilot/feasibility RCTs, as neither employed active control groups (no intervention and wait-list control) limiting inferences of mechanism of change. Lack of ITT analyses where attrition rates were high also limited inferences which was reflected in lower WoE A ratings. Methodological quality was further reduced due to lack of follow-up measurement in Kuosmanen et al. (2017), which is arguably crucial for a study evaluating SPARX as a prevention programme. Fleming et al. (2012) included a 5-week post-intervention follow-up, and therefore received higher WoE A rating than Kuosmanen et al. (2017) overall.

Finally, there was no evidence of sufficient power in the non-randomised pre-post uncontrolled trial (Lucassen et al., 2015), increasing the likelihood of false-positive inferences. Coupled with follow-up measurement that were limited to one time point (3 months post-intervention), this study achieved lowest methodological quality (WoE A) of the five studies.

Topic relevance

Primary Outcome Measures of Depression

This review favours studies which measure depressive symptoms as a primary outcome through valid, reliable and widely used measures, given that SPARX was specifically developed to target depressive symptoms.

Whilst all primary outcomes were reliable measures of depressive symptoms, receiving high WoE A ratings, several were not validated or widely used measures of depression for the specific population of participants in the sample. This is a limitation for studies that seek to address whether SPARX is effective across all educational settings.

Measures of depressive symptoms in two studies (Kuosmanen et al., 2017 & Fleming et al., 2012) were used with youth aged up to 20 years old and those excluded from mainstream school at higher risk of SEN despite not being validated for youth aged 17 years or older (Short Moods and Feelings Questionnaire) or those with additional literacy needs (Child Depression Rating Scale Revised measure).

A further study (Lucassen et al., 2015) used a depression measure (Child Depression Rating Scale Revised) for youth up to 19 years when it had only been validated for youth 12-17 years. WoE C ratings for these studies were therefore

reduced to reflect that depressive symptoms were not captured using measures that were fully validated for the specific study samples.

Intervention Delivery

This review aims to evaluate the effectiveness (as opposed to mere efficacy) of SPARX, and therefore favours studies in which SPARX was delivered with greater flexibility, i.e., limited or no monitoring/supervision during intervention completion, and in naturalistic contexts that resemble the educational settings that SPARX is likely to be implemented. WoE C for this quality dimension was therefore highest for Poppelaars et al. (2016) and Lucassen et al., (2015) because SPARX was delivered most flexibly without any supervision or monitoring of engagement, and choice of time/place in completing the game was based on student preference. However, all five studies received medium-high ratings on this dimension because SPARX was not implemented with highly structured and controlled oversight in any studies. Instead, at most SPARX was delivered with minimal supervision/monitoring during scheduled class time, completing one 20-30 minute module per week. Overall WoE C across both topic relevance quality dimensions was therefore highest for Poppelaars et al. (2016) who employed both appropriate measures for their sample and delivered SPARX flexibly.

Given that SPARX is a computerised CBT game, implementation fidelity was assumed to be consistent across all five studies included in this review and did not factor into WoE ratings. In terms of content, each variant of SPARX included within the review (SPARX, SPARX-R and Rainbow-SPARX) was implemented according to the intended programme of intervention (universal or targeted), where SPARX and Rainbow SPARX was used as a targeted intervention and SPARX-R was used as

universal intervention. For the purpose of this review, all variants and modes of delivery of SPARX were deemed of equal weighting and therefore did not factor into WoE quality ratings.

Findings

Table 7 *Study Findings*

Study	Sample size	Outcome measure	Post intervention & follow-up effect size (Cohen's *d*)	95% CI	WoE D
Fleming et al., 2012	32	Child Depression Rating Scale Revised (CDS-R)	post *d* =1.44 large 5 wk f/up – effect maintained	0.62 to 2.27*	Medium (2.25)
Kuosmanen et al., 2017	146	Short Moods and Feelings Questionnaire (SMFQ)	post *d*=0.11	-2.30 to 3.58	Medium (2.08)
Lucassen et al., 2015	21	Child Depression Rating Scale Revised (CDS-R)	post *d*=1.01 large 3m f/up – effect maintained	0.37 to 1.65*	Medium (1.89)
Perry et al., 2017	540	Major Depression Inventory (MDI)	post *d*=0.29 small 6m d=0.21 18m d=0.33	0.09 to 0.49* -0.01 to 0.42 -0.06 to 0.73	High (2.67)
Poppelaars et al., 2016	102	Reynolds Adolescent Depression Scale (RADS-2)	post *d*=0.05 3m *d*=0.17 6m *d*=0.01 12m *d*=0.41 small	-0.34 to 0.44 -0.22 to 0.56 -0.37 to 0.40 0.01 to 0.80*	High (3)

Note. CI= Confidence Interval; *WoE D*= Weight of Evidence D; *f/up*= Follow Up; *m*= months; *wk* = week; *d*= 0.2 small; *d*= 0.5 medium; *d*=0.8 large; * p<0.05; WoE Low ≤ 1.4; Medium 1.5 – 2.4; High ≥ 2.5

Table 7 presents a summary of the key findings. Effect sizes were calculated for impact of SPARX on depressive symptoms are expressed as standardised mean differences (Cohen's *d*). These were calculated from descriptive or univariate test statistics, using the Campbell Collaboration Effect Size Calculator (https://campbellcollaboration.org/research-resources/effect-size-calculator.html).

Where possible, this was calculated by the author as the difference between intervention and comparison (post-test minus pre-test) means divided by the pooled standard deviation of pre-test means. If there was insufficient data, only post-test means were used. Follow-up effect size was not able to be calculated for two studies based on insufficient data being reported in the manuscript (Fleming et al., 2012 & Lucassen et al., 2015), but are noted in the table as 'effect maintained'.

Two studies (rated medium and high quality WoE) found no significant difference at post-intervention between the SPARX and control conditions (Poppelaars et al. 2016 & Kuosmanen et al., 2017). A small post-intervention effect was found by one of the highest rated quality studies (Perry et al., 2017), however this effect was not maintained at follow up. Depressive symptoms as measures by this study were below the threshold for 'mild depression' within the SPARX group at pre- and post-intervention time points.

Two medium-rated WoE studies with the smallest sample sizes (N = 32 & 21, respectively) found significant large effects of SPARX on depressive symptoms at post-intervention, which were maintained at follow-up (Fleming et al., 2012 & Lucassen et al., 2015).

Both of these studies used a primary outcome measure (Child Depression Rating Scale Revised) where scores above the 70th percentile (raw score 30; t-score 55) are

considered the threshold for clinically significant depressive symptoms, and a decrease in CDRS-R raw score to under 30 (or a decrease of 30% or more in CDRS-R raw score) is considered sufficient for clinically meaningful change.

In the case of Fleming et al., (2012), average CDRS-R scores fell from above the clinical threshold pre-SPARX to below the threshold post-SPARX, which was maintained at 5 weeks follow up. In contrast raw scores remained above clinical threshold at pre, post and at follow-up time points in the case of Lucassen et al. (2015).

Conclusions and Recommendations

Discussion of Findings

This review evaluated how effective SPARX is at reducing depressive symptoms in young people in educational settings. Five studies met the inclusion criteria. Based on overall WoE criteria, two studies were of 'high quality' and three were of 'medium quality'. One of the 'best quality' studies showed a small effect, however it was not maintained at follow up. Two of the 'poorest quality' studies showed large effects which were maintained at follow up.

Whilst this combined data gives some support to the effectiveness of SPARX in this context, the methodological weaknesses of the studies with the largest effects leave findings inconclusive overall. More high-quality studies with a continued diverse range of education sites are needed in future trials to warrant implementation in educational settings, with particular attention to long-term follow up.

Limitations of the Review

A key limitation of this review is that the WoE ratings did not capture all possible factors of study rigour. A clear example here is that two studies had very small sample sizes (N = 32 and N=21) and yet were both rated 'medium' overall based on WoE criteria. WoE A contributed to this inflated rating in one study (Fleming et al. 2012) whereby it was rated as having 'strong evidence' (3 out of 3) for the quality dimension of sufficient power based on having technically satisfied everything within the quality rating criterion (providing a clear rationale with power calculation based on detecting a large effect). However, in practice their sample size was very small when compared to the literature as a whole, and it is unclear as to why the power calculation was based on the expectation of a large effect (Cohen's *d* >1). The coding scheme was sensitive to detecting whether sample sizes were in principle sufficiently large, but it was not sensitive enough to detect whether the sample size was sufficiently large when compared to the literature as a whole (including realistic effect size estimates based on previous studies or similar interventions).

Therefore, two of the studies that scored medium in the present review may have been more appropriately represented as 'low' given that they were both exploratory and included small sample sizes.

A further limitation of this review is that the WoE coding system did not capture cultural specificity. Given that SPARX was developed in New Zealand to be culturally specific and appropriate within New Zealand, some study locations included within this review may have found SPARX more or less inclusive of their specific culture, further impacting overall effectiveness of the intervention. This review could be improved by paying particular attention to cultural specificity. However, this factor is

not always easy to evaluate objectively based on information provided by study authors.

Despite SPARX being implemented as a prevention programme in several studies included in this review, true prevention effects were not measured within the studies included. Prevention programmes most often report on treatment effects (depressive symptoms decrease in the intervention condition compared to the control condition which remain stable) rather than true prevention effects (depressive symptoms in the control condition increase, while depressive symptoms in the intervention condition do no increase, or increase less) (Horowitz & Garber, 2006). Prevention effects may become visible if participants are followed up longitudinally, however the current review was limited by inclusion of only two studies that included longer-term follow up.

Finally, this review is limited by the subjectivity involved in appraising quantitative research without determining intra and inter-rater reliability. Whilst a published tool was used to enhance objectivity and consistency of appraisal, adaptation of the tool was influenced by the author's biases with regards to what dimensions are deemed most necessary for methodological quality, relevance, and topic relevance. Determining and reporting interrater reliability in systematic reviews is an increasingly known area of good practice and would improve future reviews in this area.

Recommendations for Future Research & Practice

The next phase of research would benefit from more trials with high methodological quality, large sample sizes, and replication of existing effects with particular consideration of long-term follow-up. Further research would benefit from

considerations of cultural specificity within educational settings when aiming to address the question of effectiveness.

Despite inconclusive findings of this review, SPARX is a pragmatic intervention with flexible delivery, particularly given the current context of online learning. SPARX therefore holds promise in bringing CBT-based intervention to a wider-reach of young people, including those in educational settings.

As a preliminary step towards dissemination, educational settings are recommended to trial SPARX with youth in their local setting. This can be supported with pre and post intervention focus groups to ascertain whether practice-based evidence further supports the emerging literature.

References

Almeida, J., Johnson, R., Corliss, H., Molnar, B., & Azrael, D. (2009). Emotional Distress Among LGBT Youth: The Influence of Perceived Discrimination Based on Sexual Orientation. *Journal of Youth and Adolescence*, *38*, 1001–1014.

Angold, A., Costello, E.J., Messer, S.C., Pickles, A.,Winder, E., Silver, D. (1995). Development of a short questionnaire for use in epidemiological studies of depression in children and adolescents: factor composition and structure across development. *Int. J. Methods Psychiatr. Res. 5*, 237–249.

Bech P, Rasmussen, N.A., Olsen, L.R., Noerholm, V, Abildgaard, W. (2001) The sensitivity and specificity of the Major Depression Inventory, using the Present State Examination as the index of diagnostic validity. *Journal of Affective Disorders, 66*(2-3),159-164.

Beck, J. S. (1995). *Cognitive therapy: Basics and beyond.* New York, NY: The Guilford Press.

Beck, A. T., Rush, A. J., Shaw, B. F., & Emery, G. (1979). *Cognitive therapy of depression*. New York: Guilford.

Beck, A. (1976). *Cognitive Therapy and the emotional disorders*. New York: Meridian.

Blakemore, S. J., Burnett, S., & Dahl, R. E. (2010). The role of puberty in the developing adolescent brain. *Human Brain Mapping*, *31*(6), 926-933.

Bohr, Y., & Merry, S. (2016). Asynchronous Etherapy for Aboriginal Communities: Experiences from Nunavut. *Journal of the American Academy of Child and Adolescent Psychiatry*, *55*(10).

Brener, N.D., Kann, L., McManus, T., Kinchen, S.A., Sundberg, E.C., Ross, J.G.(2002) Reliability of the 1999 youth risk behavior survey questionnaire. *J Adolesc Health, 31*(4), 336-342.

Calear, A. L., & Christensen, H. (2010). Systematic review of school-based prevention and early intervention programs for depression. *Journal of Adolescence*, *33*(3), 429–438.

Cheek, C., Bridgman, H., Fleming, T., Cummings, E., Ellis, L., Lucassen, M. F., Shepherd, M., & Skinner, T. (2014). Views of Young People in Rural Australia

on SPARX, a Fantasy World Developed for New Zealand Youth With Depression. *Journal of Medical Internet Research*, *16*(2)

Cheek, C., Fleming, T., Lucassen, M. F. G., Bridgman, H., Stasiak, K., Shepherd, M., & Orpin, P. (2015). Integrating Health Behavior Theory and Design Elements in Serious Games. *JMIR Mental Health, 2*(2)

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159.

Craighead, W. E., Smucker, M. R., Craighead, L. W., & Ilardi, S. S. (1998). Factor analysis of the Children's Depression Inventory in a community sample. *Psychological Assessment, 10*(2), 156-165.

Department for Education (DfE). (2020). *State of the nation 2020: children and young people's wellbeing.* Research Report. London DfE, The Stationery Office.

Ellis, L.A. (2004). Peers Helping Peers: The Effectiveness of a Peer Support Program in Enhancing Self-concept and Other Desirable Outcomes. University of Western Sydney, School of Psychology, Doctorate Thesis

Endicott, J., Nee, J., Ruoyong, Y., & Wohlberg, C. (2006). Pediatric quality of life enjoyment and satisfaction questionnaire (PQ-LES-Q): Reliability and validity. *Journal of the American Academy of Child and Adolescent Psychiatry, 45*(4), 401-407.

Fleming, T., Lucassen, M., Stasiak, K., Shepherd, M., & Merry, S. (2016). The impact and utility of computerised therapy for educationally alienated teenagers: The views of adolescents who participated in an alternative education-based trial. *Clinical Psychologist*, *20*(2), 94–102.

Fleming, T. M., Dixon, R. S., & Merry, S. N. (2012). 'It's mean!' The views of young people alienated from mainstream education on depression, help seeking and computerised therapy. *Advances in Mental Health*, *10*(2), 195–203.

Griffiths, K. M., Christensen, H., & Jorm, A. F. (2008). Predictors of depression stigma. *BMC psychiatry*, *8*(1), 1-12.

Gross, J.J., John, O.P. (2003). Individual differences in two emotion regulation processes: implications for affect, relationships, and well-being*. J. Pers. Soc. Psychol., 85*, 348–362.

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.

Horowitz, J., & Garber, J. (2006). The prevention of depressive symptoms in children and adolescents: A meta-analytic review. Journal of Consulting and Clinical Psychology, 74(3), 401-415. *Journal of Consulting and Clinical Psychology*, *74*, 401–415.

Kazdin, A. E., French, N. H., Unis, A. S., Esveldt-Dawson, K., & Sherick, R. B. (1983). Hopelessness, depression, and suicidal intent among psychiatrically disturbed inpatient children. *Journal of Consulting & Clinical Psychology, 51*(4), 504-510.

Kent, L., Vostanis, P., & Feehan, C. (1997). Detection of major and minor depression in children and adolescents: Evaluation of the mood and feelings questionnaire. *Journal of Child Psychology and Psychiatry, 38*(5), 565-573.

Kratochwill, T. R. (2003). Task Force on Evidence-Based Interventions in School Psychology. Retrieved February 14th, 2021, from http://www.indiana.edu/~ebi/documents/_workingfiles/EBImanual1.pdf

Kuosmanen, T., Fleming, T. M., Newell, J., & Barry, M. M. (2017). A pilot evaluation of the SPARX-R gaming intervention for preventing depression and improving wellbeing among adolescents in alternative education. *Internet Interventions*, *8*, 40–47.

Lucassen, M. F. G., Hatcher, S., Stasiak, K., Fleming, T., Shepherd, M., & Merry, S. N. (2013). The views of lesbian, gay and bisexual youth regarding computerised self-help for depression: An exploratory study. *Advances in Mental Health*, *12*(1), 22–33.

Lucassen, M. F. G., Merry, S. N., Hatcher, S., & Frampton, C. M. A. (2015). Rainbow SPARX: A Novel Approach to Addressing Depression in Sexual Minority Youth. *Cognitive and Behavioural Practic*e, 22(2), 203–216).

Merry, S. N., Stasiak, K., Shepherd, M., Frampton, C., Fleming, T., & Lucassen, M. F. G. (2012). The effectiveness of SPARX, a computerised self help intervention for adolescents seeking help for depression: Randomised controlled non-inferiority trial. *BMJ : British Medical Journal (Online)*, *344*.

Petticrew, M., & Roberts, H. (2003). Evidence, hierarchies, and typologies: Horses for courses. *Journal of Epidemiology and Community Health*, *57*, 527–529.

Poppelaars, M., Tak, Y. R., Lichtwarck-Aschoff, A., Engels, R. C., Lobel, A., Merry, S. N., Lucassen, M. F., & Granic, I. (2016). A randomized controlled trial

comparing two cognitive-behavioral programs for adolescent girls with subclinical depression: A school-based program (Op Volle Kracht) and a computerized program (SPARX). *Behaviour Research and Therapy*, *80*, 33.

Poznanski, E. O. and Mokros, H. B. (1996). *Children's Depression Rating Scale-Revised: manual*. Los Angeles: Western Psychological Services.

Reynolds, W. M. (2002). Reynolds adolescent depression scale: Professional manual (2nd ed.). Odessa, FL: Psychological Assessment Resources, Inc.

Rogers, R. W. (1975). A Protection Motivation Theory of Fear Appeals and Attitude Change1. *The Journal of Psychology*, *91*(1), 93–114.

Shepherd, M., Fleming, T., Lucassen, M., Stasiak, K., Lambie, I., & Merry, S. N. (2015). The Design and Relevance of a Computerized Gamified Depression Therapy Program for Indigenous Maori Adolescents. In *JMIR Serious Games*, 3(1).

Spence SH. (1998) A measure of anxiety symptoms among children. *Behav Res Ther, 36*(5), 545-566.

Spitzer, R.L., Kroenke, K., Williams, J.B.W., Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder. *Arch. Intern. Med., 166* (10),1092–1097.

Tennant, R., Hiller, L., Fishwick, R., Platt, S., Joseph., S., Weich, S. et al. (2007). The Warwick- Edinburgh Mental Well-being Scale (WEMWBS): development and UK validation. *Health Qual. Life Outcomes, 5*(1), 63.

The Children's Society. (2020). *Life on Hold Children's Well-being and COVID-19.* [Online]. London: The Children's Society.

Watanabe, N., Hunot, V., Omori, I. M., Churchill, R., & Furukawa, T. A. (2007). Psychotherapy for depression among children and adolescents: A systematic review. *Acta Psychiatrica Scandinavica, 116*, 84-95.

World Health Organisation, Mental Health Action Plan 2020, Accessed February 14th 2021 https://www.who.int/publications/i/item/9789241506021

World Health Organisation, Adolescent Health, Accessed February 14th 2021 https://www.who.int/southeastasia/health-topics/adolescent-health

Yokomitsu, K., Irie, T., Sekiguchi, M., Shimizu, A., Matsuoka, H., Merry, S. N., & Stasiak, K. (2020). Gamified Mobile Computerized Cognitive Behavioral Therapy for Japanese University Students With Depressive Symptoms: Protocol for a Randomized Controlled Trial. *JMIR Research Protocols*, 9(4).

Appendices

Appendix A – Studies excluded from the review and rationale

Table 1

Studies excluded at full paper screening stage

Study	Exclusion criteria
Eichenberg, C., & Schott, M. (2017). Serious Games for Psychotherapy: A Systematic Review. *Games for Health Journal*, 6(3), 127–135.	Not accessible (article behind a paywall)
Merry, S. N., Stasiak, K., Shepherd, M., Frampton, C., Fleming, T., & Lucassen, M. F. G. (2012). The effectiveness of SPARX, a computerised self help intervention for adolescents seeking help for depression: Randomised controlled non-inferiority trial. *BMJ : British Medical Journal*, 344.	Not in educational setting (primary care setting)
Fleming, T. M., Gillham, B., Bavin, L. M., Stasiak, K., Lewycka, S., Moore, J., Shepherd, M., & Merry, S. N. (2019). SPARX-R computerized therapy among adolescents in youth offenders' program: Step-wise cohort study. *Internet Interventions,* 18.	Not in educational setting (youth offender setting)
Perry, Y., Calear, A. L., Mackinnon, A., Batterham, P. J., Licinio, J., King, C., Thomsen, N., Scott, J., Donker, T., Merry, S., Fleming, T., Stasiak, K., Werner-Seidler, A., & Christensen, H. (2015). Trial for the Prevention of Depression (TriPoD) in final-year secondary students: Study protocol for a cluster randomised controlled trial. *Trials*, 16, 451.	Not empirical data (trial protocol)

Appendix B – Adapted Kratochwill Protocol Rationale

The coding protocol from the APA Task Force Coding Protocol by Kratochwill (2003) has been used in this review. The table below details the amendments to the protocol, together with the rationale for modifications.

Table 2

Amendments made to the Kratochwill (2003) Coding Protocol

Removed Section	Rationale
General Study Characteristics	Described in Mapping the Field
Qualitative Analysis Methods	Inclusion criteria requires quantitative data
Counterbalancing of change agent	SPARX is computerized - delivered without direct facilitation and requiring only minimal supervision from teachers when implemented within the classroom
Significance of Primary/Secondary Outcomes (D)	Outcomes are addressed in effect size table
Cultural Significance	Beyond scope of the current review
Educational/Clinical significance	Discussed within narrative of the review
Site of Implementation	Inclusion criteria require that studies are all located in educational settings
Dosage	Beyond scope of the current review
Implementation Fidelity & Characteristics of the Intervener	SPARX is computerized - delivered without direct facilitation and requiring only minimal supervision from teachers when implemented within the classroom
Cost Analysis	Beyond the scope of the current review
Replication	Beyond the scope of the current review
Sufficiently large N (adapted)	Moved into its own quality rating dimension to contribute to weight of evidence A

Appendix C – Coding protocol (completed example)

Coding Protocol: Study 4

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

Name of Coder: Date: 07/02/21

Full Study Reference in APA format:

Fleming, T., Dixon, R., Frampton, C., & Merry, S. (2012). A Pragmatic Randomized Controlled Trial of Computerized CBT (SPARX) for Symptoms of Depression among Adolescents Excluded from Mainstream Education. *Behavioural and Cognitive Psychotherapy*, *40*(5), 529–541.

Intervention Name (description of study): SPARX

Study ID Number: 4

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

Domain:
☒ School- and community-based intervention programs for social and behavioural problems
☐ Academic intervention programs
☐ Family and parent intervention programs
☐ School-wide and classroom-based programs
☐ Comprehensive and coordinated school health services

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
☒ Appropriate unit of analysis
☐ Familywise/experimenter wise error rate controlled when applicable

C. Type of Program

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

D. Stage of Program

☐ Model/demonstration programs

☒ Early stage programs
☐ Established/institutionalized programs
☐ Unknown

E. Concurrent or Historical Intervention Exposure

☐ Current exposure
☐ Prior exposure
☒ Unknown

A. 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)

C. Measurement (Estimating the quality of the measures used to establish effects)

A1 The use of the outcome measures produce reliable scores for the majority of the primary outcomes

☒Yes Well established reliability and concurrent validity (Myers & Winters, 2002)
☐No
☐Unknown/unable to code

A2 Multi-method (at least two assessment methods used)
☐ Yes
☐ No
☒ N/A
☐ Unknown/unable to code

A3 Multi-source (at least two sources used self-reports, teachers etc.)
☐ Yes
☐ No
☒ N/A
☐ Unknown/unable to code

A4 Validity of measures reported (well-known or standardized or norm-referenced are considered good, consider any cultural considerations)

☐ Yes validated with specific target group
☒ In part, validated for general population only (not validated for this specific group with lower literacy abilities however students could have them read aloud if preferred)
☐ No
☐ Unknown/unable to code

Overall Rating for measurement__3__

3= Strong Evidence **2=Promising Evidence** **1=Weak Evidence** **0=No Evidence**

B. Comparison Group

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

☐ Typical intervention (typical intervention for that setting, without additions that make up the intervention being evaluated)
- ☐ Attention placebo
- ☐ Intervention element placebo
- ☐ 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)
- ☐ Moderate (weak inference)
- ☐ High (strong inference)
- ☒ Very high (explicitly stated)
- ☐ Unable to identify comparison group

B3 Group equivalence established (select one of the following)

☒ Random assignment
☐ Posthoc matched set
☐ Statistical matching
☐ Post hoc test for group equivalence

B4 Equivalent mortality
☐ Low attrition (less than 20 % for post)
☐ Low attrition (less than 30% for follow-up)
☐ Intent to intervene analysis carried out? Findings: An intention-to-treat analysis was also undertaken, although the sample size was not adequate for this to be conducted as a main analysis.

Overall rating for Comparison group __1__

3= Strong Evidence 2=Promising Evidence 1=Weak Evidence 0=No Evidence

C. Sufficient Power

☐ Partial rationale for sample size
☒ Clearly explains rationale for sample size
☒ Sufficiently large N: Powered (80% power) to detect a large effect size (d>=1.0) with a sample size of 15/group allowing up to 50% loss of participants at follow-up.

Total size of sample (start of study): 32
Intervention group sample size (SPARX): 20
Control group sample size (Control): 12

Overall rating for Sufficient Power __3__

3= Strong Evidence 2=Promising Evidence 1=Weak Evidence 0=No Evidence

D. Follow-Up Assessment

☒ Timing of follow up assessment. Specify: 5 weeks post intervention
- ☒ Number of participants included in the follow up assessment. Specify: SPARX group n=16; Control group n=11
- ☒ Consistency of assessment method used. Specify: same measures

Overall rating for Follow-up Assessment __2__

3= Strong Evidence 2=Promising Evidence 1=Weak Evidence 0=No Evidence

Study 4: Summary of Evidence

Indicator	Overall evidence rating 0-3	Description of evidence Strong Promising Weak No/limited evidence Or Descriptive ratings
General Characteristics		
Design		Pragmatic Randomised controlled trial
Type of programme		Universal prevention programme
Stage of programme		Early stage
Concurrent/ historical intervention exposure		Unknown
Key features		
1. Measurement	3	Strong
2. Comparison group	1	Weak
3. Sufficient Power	3	Strong
4. Follow-Up	2	Promising

Appendix D – WoE criteria and ratings

Weight of Evidence A

Table 3

Criteria used when assessing Weight of Evidence A using the adapted Kratochwill (2003) Coding Protocol.

Dimension to Assess Methodological Quality (WoE A)	Summary of criteria		
	Weak Evidence (1)	**Promising Evidence (2)**	**Strong Evidence (3)**
Measurement	• Reliability coefficient ≥.5 (for primary outcome)	• Reliability coefficient ≥.7 (for primary outcome measure)	• Reliability coefficient ≥.85 (primary outcome measure)
Comparison Group	• A 'no intervention' control group was used (e.g. Waitlist or no intervention) • No evidence that groups are shown as equivalent (in terms of random assignment or equivalent attrition or using intention to treat to account for non-equivalent attrition)	• A 'no intervention' control group was used (e.g. Waitlist or no intervention) • Group equivalence by random assignment • Shown that there is low attrition or intention to treat if high attrition	• Active Control Group used • Group equivalence by random assignment • Shown that there is low attrition or intention to treat if high attrition
Sufficient Power	• Partial rationale for sample size but without any formal sample size calculations	• Clear rationale for sample size but without any formal sample size calculations	• Clear rationale with power calculation or equivalent sample size calculation
Follow-Up	• Follow-up assessments conducted at least once with only some of the original participants invited	• Follow-up assessments conducted at least once, with the majority of the original participants invited and using similar measures to analyse primary outcome	• Follow-up assessments conducted over multiple intervals, inviting all the original participants and using similar measures to analyse the data for primary outcome

Table 4

WoE A Ratings

Study	Quality ratings assigned for the 4 dimensions				Overall WoE A (mean score to 2 decimal points)
	Measurement (0-3)	Comparison Group (0-3)	Sufficient Power (0-3)	Follow-Up (0-3)	
Perry et al., 2017	3	3	3	3	3
Poppelaars et al., 2016	3	3	3	3	3
Kuosmanen et al., 2017	3	1	3	0	1.75
Fleming et al., 2012	3	1	3	2	2.25
Lucassen et al., 2015	3	N/A	0	2	1.67

Weight of Evidence B

This criteria is derived from evidence typologies that suggest which designs are more appropriate for certain review questions (Petticrew & Roberts, 2003). For this review, WoE B considered the relevance of the methodology for evaluating the effectiveness of an intervention and the author selected design and generalisability as two key factors relating to effectiveness.

Table 5

WoE B Criteria

Criteria	Ratings	Rationale
Study Design	3.Definitive RCT 2. Pilot/Feasibility RCT 1.Non-randomised study (Cohort study, case-control studies, cross sectional survey, case reports)	Effectiveness relies on being able to conduct formal hypothesis testing for outcome measures and requires causality inferences to be made.

Generalizability of Study Sample	3. Multi-site, wide geographic coverage (e.g. across a country or multiple countries) 2. Multi-site, narrow geographic coverage (e.g. within a city) 1. Single-site (e.g. individual school)	Effectiveness relies on being able to generalize results to diverse student populations

Table 6

WoE B Ratings

Study	Quality ratings assigned for the 2 dimensions		Overall WoE B
	Study Design (1-3)	Generalizability of study sample (1-3)	
Perry et al., 2017	3	2	2.5
Poppelaars et al., 2016	3	3	3
Kuosmanen et al., 2017	2	3	2.5
Fleming et al., 2012	2	3	2.5
Lucassen et al., 2015	1	2	1.5

Weight of Evidence C

This criteria is derived from Harden & Gough (2012) to discern topic relevance to the specific review question.

Table 7

WoE C Criteria

Criteria	Ratings	Rationale
Primary outcome measure of depressive symptoms	3. Primary outcomes include a valid, reliable and widely-used measure of depression symptoms for this specific population. 2. Primary outcomes include a non-validated or not widely used measure of depression for this specific population. 1. Depression is not included as a primary outcome (but is a secondary or exploratory outcome)	The review is concerned with depressive symptoms
Format of Delivery	3. Intervention delivered with flexibility (no structured monitoring or supervision of engagement) 2. A minimum threshold of engagement was enforced (some oversight and monitoring of engagement) 1. Highly structured and controlled oversight ensuring engagement	The aim of the review is to report on effectiveness rather than efficacy, therefore the most relevant studies are those where the intervention was delivered flexibly and compliance was not enforced ('real world' context / naturalistic)

Table 8

WoE C Ratings

Study	Quality ratings assigned for the 2 dimensions		Overall WoE C
	Primary Outcome Measure of Depression (1-3)	Format of Delivery (1-3)	
Perry et al., 2017	3	2	2.5
Poppelaars et al., 2016	3	3	3
Kuosmanen et al., 2017	2	2	2
Fleming et al., 2012	2	2	2
Lucassen et al., 2015	2	3	2.5

Table 9

WoE D Ratings

Study	WoE A: Methodological Quality	WoE B: Methodological Relevance	WoE C: Topic Relevance	WoE D: Overall weight of evidence (average score of A,B and C)
Perry et al., 2017	High 3	High 2.5	High 2.5	High 2.67
Poppelaars et al., 2016	High 3	High 3	High 3	High 3
Kuosmanen et al., 2017	Medium 1.75	High 2.5	Medium 2	Medium 2.08
Fleming et al., 2012	Medium 2.25	High 2.5	Medium 2	Medium 2.25

Lucassen et al., 2015	Medium 1.67	Medium 1.5	High 2.5	Medium 1.89

Table 10

WoE A,B & C Qualitative Descriptors

Descriptor	Average Score Band
High	≥ 2.5
Medium	1.5 – 2.4
Low	≤ 1.4