Case study 1: An Evidence-based practice review report.

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

Are Nurture Groups (NGs) effective in improving the wellbeing of students with Social, Emotional and Mental Health (SEMH) needs?

Summary

Children with Social, Emotional and Mental Health (SEMH) needs are more likely to have lower academic and wellbeing outcomes. This often continues into their adult life effecting their opportunities and relationships. Educational psychologist, Marjorie Boxall, developed the Nurture Groups (NGs) intervention to improve the wellbeing of children who presented with challenging behaviour due to their early experiences. NGs utilised an understanding of attachment theory and Maslow's (1970) hierarchy of needs, to create a space within school where positive attachments could be fostered. In this review, the effectiveness of NGs will be evaluated through studies published following Hughes and Schlosser's (2014) review. Six papers were included based on their criteria, one of which addressed the secondary review question of the successful strategies used in NG. Further evidence of the effectiveness of NGs at reducing SEMH needs in children was found. However, the need for further studies that include control groups and sample sizes that fulfil the power calculations is discussed.

Introduction

The need for interventions to be evidence based is apparent, especially within the context of implementation within education settings. Ever-increasing demands on school staff workloads coupled with the diminishing relative funding, schools must commit to a proven and effective

intervention. Students with Social, Emotional and Mental Health (SEMH) needs arguably have the greatest need for intervention that is evidence based to improve their outcomes. Their vulnerabilities extend to both their academic and social lives, self-concepts and outcomes. In childhood, their academic disadvantage comes from limited time spent in lessons, internalising and externalising behaviour problems, all having the potential in later life to culminate into lower financial stability, criminality and even psychiatric hospitalisation (Cannon et al., 2013). Socially, they are also far less likely to maintain relationships, both with peers and adults, resulting in detrimental emotional outcomes that are likely to follow them into adult life (Cannon et al., 2013).

In light of the COVID-19 pandemic, data from NHS Digital (2020) reveals that children and young people's mental health has been impacted negatively. The impetus therefore is for schools to provide access to effective interventions that improve outcomes for children who have SEMH needs. An Ofsted (2011) survey promoted the potential for NGs to improve their social and emotional development. Furthermore, a review conducted by Hughes and Schlosser (2014) highlighted evidence for the effectiveness of NGs specifically for children with SEMH needs. In addition, they revealed the benefits to those children within schools that adopted NGs, even though they did not take part in the groups themselves.

Nurture Groups

In 2019, Nurture UK celebrated 50 years since the NG provision was first introduced by Marjorie Boxall (Boxall, 2002; Nurture UK, 2019). Initially, NGs aimed to address the growing population of children entering primary school displaying unmanageable behaviour (Nurture UK, 2019). For example, the initial NGs included students from what is now known as the "Windrush Generation", migrant children who had experienced trauma of migration and low socio-economic standards (Boxall, 2002).

The 'classic' NG is designed to facilitate 10-12 children (aged 4-8) within a safe space, accompanied by a teacher and teaching assistant, once every school day for 2.5 hours each

(Boxall, 2002). Adaptions to this model have included inclusion of children up to the age of 15, utilising age differentiated topics, as well as part time models that involve less and shorter sessions (Cooke et al., 2008). Although students attend their mainstream classes alongside the NG provision, the intervention aims for full transition gradually, to the mainstream timetable (Seth-Smith et al., 2010).

Psychological Theory

The theorised cause of the unmanageable behaviour was poor early attachment, therefore concepts from Bowlby's (1969) attachment theory underpin NGs. Utilising the understanding that children have an innate need to form attachments, that they then use to develop an internal working model, NGs offer children reparative attachments within a familiar and consistent setting, namely school (Boxall, 2002). The safe space that NGs offer provides a secure base for the child to explore their understanding of the world, presumed initially negative, and grow in their ability to understand and regulate emotions as well as learn social and academic skills (Hughes & Schlosser, 2014).

Another contributing theory to the efficacy of NGs is Maslow's (1970) hierarchy of need. The concept that all the aspects of the child's needs are met within NGs is given weight by the initial provision of a safe and reliable space. Within which, a sense of belonging is instilled and students take part in activities partly designed to boost their self-esteem. Maslow (1970) suggests these elements are essential to inspire motivation to learn, arguably offering an insight into the emotional and behavioural difficulties displayed by vulnerable children. Furthermore, Cameron and Maginn (2011) champion the importance of within school relationships, mainly due to the amount of time and opportunity for social experience offered. Their value is emphasised in COVID's effect on school attendance and the rise in emotional and mental health issues (NHS Digital, 2020).

Rationale and relevance to educational psychology practice

SEMH needs constitute a Special Educational Need (SEN) and therefore students who have them are supported within the SEN Code of Practice (DfE & DoH, 2014). This highlights the importance of the wishes of the parents and children, particularly for mainstream schools needing to ensure that students with SEN are not disadvantaged. When we consider that significantly impacting the learning of other students is the other element, alongside parental and child wishes, that can prevent a child being educated in mainstream, the impetus to address SEMH needs is clear.

Furthermore, government guidance for promoting positive mental health and behaviour recommends, among other interventions, NGs (DfE, 2018). As champions for the best interest of the child and activists in removing barriers to their wishes, Educational Psychologists (EPs) need to be able to advise schools to implement effective interventions. NGs have a sound psychological basis, however reviews into current research is essential to maintain evidence of efficacy and therefore quality of EP advice.

Review question

Hughes and Schlosser (2014) systematically reviewed research on NGs and found benefits to student wellbeing. Despite this, the Early Intervention Foundation (2017) still only rate the evidence for NGs as “preliminary”. Furthermore, guidance from the DfE and DoH (2014) gave greater impedance for the child’s right to inclusion in mainstream in spite of any SEMH needs they may be experiencing. Akin to Hughes and Schlosser (2014) this systematic review will consider these questions:

- Are NGs effective in improving the well-being of students with SEMH needs?
- If so, what were the strategies adopted by the facilitator or setting which appeared to be effective?

Critical review of the evidence base

Literature search

The following three databases were searched for this review; PsychINFO, Web of Science and Education Resources Information Centre (ERIC). Following on from the work of Hughes and Schlosser (2014) the searches were restricted to peer reviewed journal articles published either in or since 2014.

The search term used was “nurture group*” (*= truncation) as the intervention does not have any synonyms.

Inclusion and exclusion criteria

Studies were selected in line with the inclusion criteria set out by Hughes and Schlosser (2014), with the date added, see Table 1.

Table 1 – Inclusion and exclusion criteria:

	Inclusion Criteria	Exclusion Criteria	Rationale
1	Peer-reviewed journal articles.	Non-peer reviewed journal articles, such as books, dissertations or reviews	Peer-reviewed journal articles are screened for validity and quality prior to publishing and therefore are likely to be of higher quality.
2	Nurture group intervention (classic or adapted).	Any other intervention, even if Nurture based.	Nurture UK (2019) outlines specific criteria for groups to constitute a NG. Therefore, it is

			important that this review focuses on the intended intervention.
3	Participants were school-aged (4–18 years old).	Participants outside of that age range.	Although EPs work with young people up until the age of 25, NGs are specifically targeted towards children within school settings and therefore nurture interventions aimed at ages outside of these areas should not be considered.
4	Quantitative studies, focusing on the effectiveness of nurture groups OR observational studies examining particular strategies used by NG teachers.	Purely qualitative studies	In line with the hierarchy of studies most appropriate to address effectiveness (Petticrew & Roberts, 2003), purely qualitative studies lack the generalisability to fully appraise an intervention.
5	A minimum of one outcome measure	No outcome measures that relate to SEMH needs	This review is directly targeting NGs

	attributed to SEMH needs in quantitative studies.		effectiveness for supporting children with SEMH needs. Therefore, the evidence needs to clearly measure the effect in that area.
6	Published in or post 2014.	Published prior to 2014	Studies published prior to 2014 were considered within Hughes and Schlosser's (2014) previous review.

List of included studies

Studies included are show in Table 2, a detailed "mapping of the field" can be found in appendix A. For excluded studies see appendix B.

Table 2 – reference list of included studies:

	Included studies
1	Cubeddu, D. and MacKay, T. (2017) The attunement principles: a comparison of nurture group and mainstream settings. *Emotional and Behavioural Difficulties*. *22*(3), 261-274. https://doi.org/10.1080/13632752.2017.1331985
2	Cunningham, L., Hartwell, B. and Kreppner, J. (2019) Exploring the impact of Nurture Groups on children's social skills: a mixed-methods approach. *Educational Psychology in Practice*. *35*(4). 368-383. https://doi.org/10.1080/02667363.2019.1615868

3 Grantham, R and Primrose, F (2017) Investigating the fidelity and effectiveness of Nurture Groups in the secondary school context. *Emotional and Behavioural Difficulties. 22*(3), 219-236. https://doi.org/10.1080/13632752.2017.1331986

4 Hibbin, R and Warin, J. (2016). Nurture groups in practice: children; classes; schools: Final report of Comparative study of nurture groups and alternative provisions for children with social, emotional and behavioural difficulties. *Nurture Group Network.* https://doi.org/10.13140/RG.2.1.3477.1609

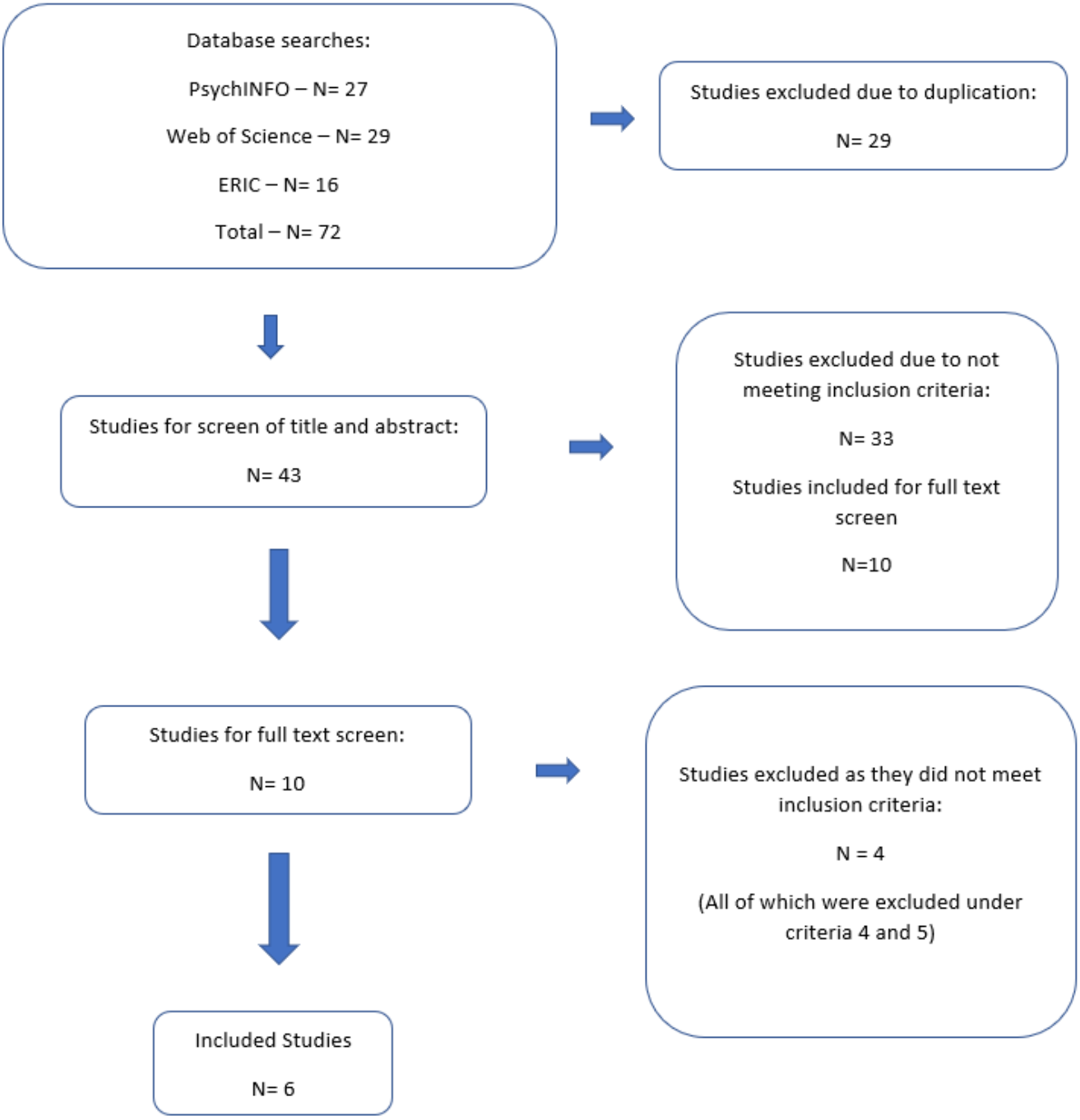
5 Lyon, L. (2017). A pilot study of the effectiveness of a nurture group in a secondary special school*. International Journal of Nurture in Education*, *3*(1), 6-17. https://www.nurtureuk.org/sites/default/files/lyon.pdf

6 Sloan, S, Winter, K, Connolly, P, and Gildea, A. (2020). The effectiveness of Nurture Groups in improving outcomes for young children with social, emotional and behavioural difficulties in primary schools: An evaluation of Nurture Group provision in Northern Ireland. *Children and Youth Services Review*, *108*, 104619. https://doi.org/10.1016/j.childyouth.2019.104619

Flowchart of search and screening process

The below flowchart (figure 1) visualises the search, screening and selection process for the studies to be included within this review. Please see Appendix B for details on the excluded studies and the criteria used to exclude them.

Figure 1 – Flowchart of search and screening process, number of studies found, excluded then finally included.



Weight of Evidence (WoE)

Gough's (2007) framework has been utilised to appraise the six studies in this review. It consists of three main dimensions (WoE A, B and C) and one overall rating (WoE D). WoE A evaluates the quality of the study, in this review the Downs and Black checklist (1998) was used in line with Hughes and Schlosser (2014), see appendix C. WoE B focuses on the design of the study and its relevance to the research question, see appendix D. WoE C is used to evaluate the topic

relevance of the study, see appendix E. WoE D represents an average of A, B and C and gives an overall rating, see table 3.

Table 3 – Overall WoE scores

Study	WoE A	WoE B	WoE C	WoE D (Average)
Cubeddu and Mackay (2017)	1.8 Medium	2 Medium	1.5 Low	1.7 Medium
Cunningham et al. (2019)	2.1 Medium	2 Medium	2.3 Medium	2.1 Medium
Grantham and Primrose, (2017)	1.9 Medium	2 Medium	2 Medium	1.9 Medium
Hibbin and Warin, (2016)	0.9 Low	2 Medium	1 Low	1.3 Low
Lyon (2017)	1.1 Low	2 Medium	1.5 Low	1.5 Low
Sloan et al. (2020)	3 High	3 High	2.8 High	2.9 High

Qualitative descriptor criteria: 0.9 -1.6 = “low”, 1.7- 2.3 = “medium” and 2.4 - 3 = “high”.

Participants

Across the five studies addressing the primary review question there were 442 participants, it is worth noting Sloan et al. (2020) represents 87% of participants. Their ages ranged from 6 – 13 years old, therefore representing both primary and secondary settings. The total number of schools included was 67 schools, Sloan et al. (2020) represents 66% of schools. All of the studies were conducted within the United Kingdom which suggests generalisability to countries with similar education systems. Gender was not reported by Grantham and Primrose (2017) to

differentiate their 24 participants. In fact, no discernible data was reported to differentiate the participants, despite this impacting their WoE C in part, they still scored high due to particular strengths in objectivity and targeting of SEMH needs through their outcome measures. Excluding their study, 65% of the total participants were male, this may limit the extent to which findings can be generalised to the female population. Furthermore, only two studies contributed other demographic data, firstly Cunningham et al. (2019) reported 15 out of 16 participants were White British with the remaining being Asian. Secondly, Sloan et al. (2020) accounted for greater detail for both their intervention and control groups, including ethnicity, looked after children (LAC), social services involvement, SEN and eligibility for Free School Meals (FSM). This translated into their comparatively higher ratings in WoE C which accessed, in part, their consideration of confounding variables (Table E1). To a lesser extent their consideration of their participants characteristics also resulted in a higher WoE A rating due to the Downs and Black Checklist (1988) specifying this in their "reporting" section. Additionally, Lyon (2017) informs us that their four participants all attend a specialist setting due to experiencing SEMH needs.

As the study by Cubeddu and Mackay (2017) was the sole study addressing the secondary review question it shall be addressed separately in each section. They undertook observations of five teachers, one of whom was a NG teacher in a special school and the others were mainstream primary teachers. All of the teachers were female and taught classes of children aged 5-6. As no other demographic information was gathered this impacted their WoE C participants section negatively.

Type of Nurture Group

The "classic" NG model was utilised in three of the studies (Cubeddu & Mackay, 2017; Grantham & Primrose, 2017; Lyon, 2017). However, Grantham and Primrose (2017) did use adapted activities. Cunningham et al. (2019) also studied an adapted NG that was part time. Sloan et al. (2020) included both "classic" and "part time" models in their research and compared their efficacy. While the type of NG studied did not impact their WoE ratings directly, the quality of their

reporting was accessed through WoE A. Their WoE A score was impacted by their reporting of their method in relation to types of NG studied. For example, Hibben and Warin (2016) included three different NG conditions, however did not report the distribution of this variance in their outcome measures. Therefore, despite Sloan et al. (2020) also reporting on more than one type of NG, their WoE A rating was far superior due in part to the quality of reporting their method.

Research Design

Four of the five studies that relate to the primary review question utilised quasi-experimental designs. They also utilised a mixed methods approach, though the qualitative data was not accessed as part of this review. Sloan et al.'s (2020) study was a non-randomised control trial and rightfully received a higher WoE B rating than the other studies that all received a lower rating due to their homogeneity of research design. All five studies used pre and post scores.

Cubeddu and Mackay (2017) used observation in the form of an event sampling design. Due to the flexible nature of, and the adaptations made to, the Downs and Black Checklist (1998), despite the variation of design and outcome measure, the six studies could all be fairly compared within the WoE A rating, see appendix C.

Measures

As in Hughes and Schlosser's (2014) review, the Boxall Profile (BP) was the most commonly used measure, with four of the five studies relating to the first review question utilising it. The BP is completed by the teacher of the child and consists of two sections, the developmental strands (measuring social, emotional and mental health progress as well as internalisation of controls) and the diagnostic profile (measuring behaviours that inhibit the child's involvement with school) (Nurture UK, 2019). It is worth noting the BP for young people (BPYP) is the officially adapted measure for secondary school children, utilised by Grantham and Primrose (2017).

The BP has been shown by Croft et al. (2015) to have a high concurrent validity with the Strengths and Difficulties Questionnaire (SDQ). The SDQ (Goodman, 1997) also being a teacher completed instrument that measures behavioural functioning in children in two areas, main total difficulties and prosocial behaviours. The only study to utilise both the BP and SDQ was Sloan et al. (2020), this is reflected in their WoE C rating for outcome measure relevance. The two other studies that scored well within that WoE C subscale were Grantham and Primrose (2017), as mentioned above, and Lyon (2017). The latter of whom, in addition to BP, utilised the Pupil, Attitude to Self and School (PASS; RAND, 2020) and structured observations using interval recording of observed on task behaviour. The PASS on the one hand has not had its reliability or validity tested within a peer reviewed article, however has been standardised amongst a sample of more than 600,000 children and shows strong face validity in the questions that it includes and their synchronicity with SEMH needs (CORC, 2020). Conversely, Hibbin and Warin (2016) only used the BP's diagnostic section, which was not fully explored in their analysis, resulting in the lowest WoE C rating.

Cunningham et al. (2019) used two different outcome measures focused on the social aspect of wellbeing; the Child Role Play Measure (CRPM) and the Taxonomy of Problematic Social Situations (TOPSS). The CRPM (Dodge et al., 1985) was completed by the researcher with the child who speculates how they would react to examples of social situations. The TOPSS (Nangle et al., 1994) is completed by the teacher and relies on their assessment of the individual child's managing of social situations.

Sloan et al. (2020) and Cunningham et al. (2019) were the only studies that reported the internal consistency of their outcome measures, both in the form of Cronbach's Alpha (*a*) and therefore received the relevant merit in the reliability and validity section of the WoE C. Sloan et al. (2020) showed a high level of consistency for both the BP and SDQ (all exceeding *a*=0.798), but also showed a high consistency for their bespoke self-report questionnaire for a child's enjoyment of school (*a*=0.799). Cunningham et al. (2019) however measured their internal consistency pre and post test for the CRPM and TOPSS. While the TOPSS level of consistency was high both pre and

post (*a*=0.97 and *a*=0.98 respectively), the CRPM showed low consistency in the pre-condition (*a*=0.47). This was attributed to the children's hypothetical responses being inconsistent and varied due to disorganised mental representations of their own reactions. Furthermore, even the high reliability seen in the post measure (*a*=0.79) does not account for the inherent unreliability of a child's imagination of how they might react, compared with an observational measure. Their detailed reporting of this and consideration throughout their 'results' section contributed to their strong WoE C rating.

Cubeddu and Mackay (2017) were the two observers, using event sampling to compare the number of attunement principles demonstrated over a 60-minute observation by a NG teacher compared to mainstream peers. The inter-observer reliability was measured using two 12-minuite sampling periods (one NG and one mainstream), which resulted in a Pearson correlation of r=0.954. Although this may indicate a strength in the congeniality between observers, the internal consistency of the behaviour amounting to attunement was not tested. In addition, the sample was only 20% of the length of the real-world observation. Therefore, reliability diminishing over duration, especially considering the element of fatigue, was not considered. This apparent oversight impacted upon their low score for their in the outcome measures section of the WoE C rating.

Findings

All studies found that NGs had a positive effect on outcome measures, see Table 6. However only three studies reported effect sizes which resulted in higher WoE A and C ratings. These were Cunningham et al. (2019), Sloan et al. (2020) and Grantham and Primrose (2017), however they each used different calculation methods for effect sizes (Cohen's d, Hedges g and Eta Squared (η2) respectively). Glen (2016), states Cohen's d and Hedges g are comparable, so long as g is used in sample sizes over 20, which Sloan et al. (2020) have done, see Table 4. However, η2 is

not, and although it can be categorised by benchmarks, the effect sizes have been converted to Cohen's d using the Campbell Calculator in Table 5 to allow for accurate comparison.

Table 4 – Effect size measurement descriptor comparison table (Cunningham et al., 2019; Glen, 2016; Grantham and Primrose, 2017; Sloan et al., 2020):

Effect Size Benchmarks

Statistic	Small	Medium	Large
Cohen's *d*	0.2	0.5	0.8
Hedges g	0.2	0.5	0.8
Eta Squared (η^2)	0.01	0.06	0.14

Table 5 – BP pre/post t-test results (Grantham & Primrose, 2017):

Developmental strand	**Presentation of analysis (two tailed)**	**Significant**	**Effect size (η2)**	**Effect size (*d*)**	**Confidence Intervals (95%)**
Purposeful attention (A)	*t(24) = −3.961, p < .001*	*Yes*	*.4*	*-1.14*	*-1.75, -0.53*
Constructive participation (B)	*t(24) = −3.620, p < .001*	*Yes*	*.4*	*-1.05*	*-1.65, -0.44*
Connects experiences (C)	*t(24) = −4.172, p < .000*	*Yes*	*.4*	*-1.20*	*-1.82, -0.59*
Insightful involvement (D)	*t(24) = −4.818, p < .000*	*Yes*	*.5*	*-1.39*	*-2.02, -0.76*
Cognitive engagement (E)	*t(24) = −4.079, p < .000*	*Yes*	*.4*	*-1.18*	*-1.79, -0.56*
Accommodates to others (F)	*t(24) = −3.630, p < .001*	*Yes*	*.4*	*-1.05*	*-1.65, -0.44*
Constructive responses (G)	*t(24) = −5.992, p < .000*	*Yes*	*.6*	*-1.73*	*-2.39, -1.07*
Maintains standards (H)	*t(24) = −4.226, p < .000*	*Yes*	*.4*	*-1.22*	*-1.84, -0.60*
Emotionally secure (I)	*t(24) = −2.909, p < .008*	*No*	*n/a*	*-0.84*	*-1.43, -0.25*

Accepts constraints (J)	*t(24) = −2.357, p < .027*	*No*	*n/a*	*-0.68*	*-1.26, -0.09*

Diagnostic profile	***Presentation of analysis (two tailed)***	**Significant**	**Effect size**	**Effect size (*d*)**	**Confidence Intervals (95%)**
Disengaged (Q)	*t(24) = 3.406, p < .002*	Yes	.3	0.98	0.38, 1.58
Self-negating (R)	*t(24) = .891, p < .382*	No	n/a	0.26	-0.31, 0.83
Undifferentiated attachments (S)	*t(24) = 1.661, p < .110*	No	n/a	0.48	-0.09, 1.05
Inconsequential behaviour (T)	*t(24) = 3.041, p < .006*	No	n/a	0.88	0.29, 1.47
Craves attachment (U)	*t(24) = 1.518, p < .143*	No	n/a	0.44	-0.13, 1.01
Avoids/rejects attachment (V)	*t(24) = 1.773, p < .089*	No	n/a	0.51	-0.06, 1.09
Insecure sense of self (W)	*t(24) = 2.491, p < .020*	No	n/a	0.72	0.14, 1.30
Negative towards self (X)	*t(24) = .911, p < .372*	No	n/a	0.26	-0.31, 0.83
Negative towards others (Y)	*t(24) = .194, p < .848*	No	n/a	0.06	-0.51, 0.62
Wants, grabs, disregards others (Z)	*t(24) = 2.983, p < .007*	No	n/a	0.86	0.27, 1.45

Although both Grantham and Primrose (2017) and Cunningham et al. (2019) measured effect size within participants pre and post intervention, they utilised different outcome measures. The latter found a near significant and medium effect (p=0.055, d=0.52) through the teacher reported TOPSS and a significant and large effect (p=0.02, d=0.97) through the child reported CRPM. The TOPSS' Reliability Change index was reported, due to its high reliability, showing 56% (N=9) experienced a positive change and 25% (N=4) no change with the remaining 19% (N=3) experiencing regression in social skills. This suggests an improvement in social skills caused by NGs, both self and teacher reported, can be tentatively supported due to the medium overall WoE D rating obtained by the study.

Grantham and Primrose (2017) also received a medium WoE D rating, despite having made an error in t test reporting by inverting the negative and positive values (e.g. developmental scores increased post intervention despite table 4's appearance). As the type 1 error was observable and corrected for in my analysis, accurate comparison is possible. Their findings suggest that the improvements in eight of the ten developmental strands were significant and of a large effect size, whereas in the diagnostic section this was only true for the first strand. However, all BP strands moved in the direction evidencing a positive effect of the NG on the child's SEMH needs. The most significant gains were to the children's interaction within class and making of friends as well as a significant increase in the strand indicative of resilience. Their overall medium WoE D rating indicates the due weight that can be given to their findings, especially when considering that the effect sizes, within the developmental strands specifically, were large.

Reporting of effect sizes for the pre and post measures for the intervention group in Sloan et al. (2020) allows for comparison with Grantham and Primrose (2017), however they were generalised for the developmental strand ($p<0.001$, $d=1.82$) and diagnostic profile ($p<0.001$, $d=-1.13$) overall via averaging. This may be viewed as a weakness of the study as it offers less detail, however in relation to comparing their findings with Grantham and Primrose (2017), the consistence of significant and very large effect sizes for developmental strands is clear. Although the diagnostic profiles post intervention lower for both, the significance and effect sizes are inconsistent.

Considering effect sizes were only reported in half of the studies included in this review a direct comparison between study's findings was not possible. However, a summary of the findings from all six studies can be found in table 6 below.

Table 6 – Overall findings comparison table

Study	Number of Participants	WoE D Rating	Summary of findings
Cubeddu and Mackay (2017)	N=5 (teachers)	1.7 (Medium)	Mainstream teachers differed significantly between each other in use of attunement principles ($p<0.001$, $\chi2=56.422$) Overall the NG teacher used significantly more attunement principles than their mainstream colleagues ($p<0.001$, $\chi2=219.676$), this equated to between 75% and 300% more behaviours observed.
Cunningham et al. (2019)	N=16	2.1 (Medium)	TOPSS: pre/post change was nearly significant ($p=0.055$, $d=0.52$), RCI* showed 56% positive change, 25% no change and 19% negative change. CRPM: pre/post change was significant with a large effect size ($p=0.002$, $d=0.97$)
Grantham and Primrose (2017)	N=24	1.9 (Medium)	BP developmental strands: all improved post intervention, however only 8/10 significantly with large effect sizes ($p<0.001$, $d=-1.05$~-1.73**) BP diagnostic profiles: all improved post intervention however only 1/10 (“Disengagement”) significantly with a large effect size ($p<0.002$, $d=0.98$, 95% CI [0.38,1.58])
Hibbin and Warin, (2016)	N=14	1.3 (Low)	Only BP diagnostic profile overall scores obtained for 12/14 participants. Of these 10 showed improvement. Average change in diagnostic profile score post intervention was -20.58. Data provided was insufficient for statistical analysis.

Lyon (2017)	N=4	1.5 (Low)	Correlation between time in the NG and improved on-task behaviour was positive for 2/4 participants and negative for the other two. All four participants showed improvement in BP scores post intervention. PASS: All four participants showed improvement in most areas. The exception was one participant showed a decline in “Self-regard, as a learner”, “General work ethic” and “Response to curriculum demands”. Data provided was insufficient for statistical analysis.
Sloan et al. (2020)	N=384 (intervention = 296, Control = 88)	2.9 (High)	Post intervention improvement in BP developmental strand in intervention group ($p<0.001$, $d=1.817$) compared to control ($p=0.686$, $d=-0.031$). Post intervention improvement in BP diagnostic profile in intervention group ($p<0.001$, $d=-1.128$) compared to control ($p=.746$, $d=-0.023$). Post-test means to show the difference in outcome of intervention over control showed significance and large effect size in BP developmental strand ($p<0.001$, $g=1.352$, 95% CI [0.098,1.728]) and diagnostic profile ($p<0.001$, $g=0.904$, 95% CI [-1.251, -0.557]) SDQ separate sections and total difficulties improved significantly with medium to large effect sizes post intervention improvement in intervention group ($p<0.001$, $d=-1.622$~1.008) compared to control ($p=0.003$~0.815, $d=-0.23$~0.158).

Post-test means to show the difference in outcome of intervention over control showed significance and large effect size in all areas of the SDQ. The largest effect size was for total difficulties ($p<0.001$, $g=-1.303$, 95% CI [-1.696, -0.909]).

Of the academic outcomes all improved on average post-test, however only “Enjoyment of School” improved significantly and with a medium effect size in the intervention group when compared to control ($p=0.002$, $g=0.528$, 95% CI [0.199,0.857]).

In addition, participants in larger schools showed a higher level of SDQ “peer problems” post-test ($p<0.001$). As well as participants having a lower baseline score being significantly correlated with a greater amount of progress amongst most outcome measures. For example, BP developmental strand ($p<0.001$) and SDQ conduct problems ($p<0.001$).

---

*Reliability Change Index – measures the reliability of individual improvement seen, particularly appropriate for studies with a small sample size (Jacobson & Truax, 1991).

**For more detail please see Table 5.

As detailed in Table 6, Hibbin and Warin (2016) only reported the diagnostic profile overall scores, however they did gather them over three separate time points, although these were inconsistently distanced between participants. Their analysis of their data is minimal and, due to its nature, the findings only show the diagnostic profile scores improving over time within NGs for ten of the twelve participants. However, it is apparent from the WoE ratings that their study is of low quality and suffers from many limitations, chief among which being their WoE A. Additionally, their WoE C included the observation that their objectivity was questionable as their study had been commissioned and supported by what is now Nurture UK. Therefore, their study's findings should be given the least weight.

The other study to obtain an overall low WoE rating was Lyon (2017), although in relation to their BP data they provided an in-depth record of the four participants individual item scores across the three equal time points. They found that all four students made progress with their social emotional development over time, but interestingly they all showed the least progress in their removal of barriers to learning. In addition to the BP, they measured PASS and on task behaviour. PASS data showed that three out of the four participants showed increased positivity in attitudes towards themselves and school, however the other participant either maintained or decreased in three of these attitudes. They suggest this may be indicative of adolescent behaviour, although their observation data also shows that two of the four participants saw a reduction in their on-task behaviours with the NG settings, neither of these individuals was the participant who showed declining PASS scores. As the study with the smallest sample size and by not exploring any demographic data that may have contributed to the change, their findings will not carry sufficient weight. This is solidified by their WoE A and C ratings judging their lack of power and consideration of variables within participants respectively.

Conversely the strongest study included in this review has more than sufficient power and multiple outcomes measures that have high internal consistency and concept validity (BP and SDQ). Additionally, Sloan et al. (2020) was the only study to involve a control group, shown in their WoE

B rating, which therefore reduces the likelihood of inflated effect sizes (Glen, 2016). As such there were no statistically significant or large effect sizes seen in the pre and post intervention means for the BP or SDQ in the control group. This is in contrast to all of the BP sections and all but one of the SDQ sections being statistically significant and having a large effect, the exception was reduction in conduct problems which only had a medium effect size (d=-0.681).

Further analysis by Sloan et al. (2020) between the post scores for both the intervention and control groups revealed that the attendance of NGs produced large effects for the developmental strand (g=1.352) and the diagnostic profile (g=-0.904) of the BP as well as the total difficulties score (g=-1.303) and prosocial scores (g=0.926) of the SDQ. Another key strength of the study was their consideration of potential significance of many confounding variables represented in their WoE C rating. No statistically significant change in the benefits of NGs were found in gender, children with EAL, level of social care involvement (including LAC), FSM eligibility, SEN or receiving full time or part time NG provision. The two variables that were found to have a significant impact were that of the size of the school and the individual's baseline score. It was found that students in larger schools made less progress as measured by the BP as well as maintained higher peer problems as measured by the SDQ. In addition, it was found that students with lower baseline scores in the BP and SDQ saw more significant progress post intervention. Though not evident within Sloan et al.'s (2020) sample, as there was no significant change based on belonging to a vulnerable group, this may be pertinent as lower baseline scores are associated with children who are LAC (Sloan et al., 2020). However, an alternative explanation for the greatest improvement being seen amongst children who initially scored low would be the tendency for significantly low or high scores to naturally become more moderate simply through repeated test, e.g. regression toward the mean.

Finally, Cubeddo and MacKay (2017), as the only study to meet the criteria of the second review question, analysed their event sampling using the Chi-square goodness-of-fit test ($\chi 2$). This test is limited to showing the independence of two variables and their interactions difference from

hypothesised observations. Their findings showed that the NG teacher used significantly more attunement behaviours than her mainstream peers (between 75-300%). Crucially, they found the most lacking area of attunement in mainstream was that of deepening discussion, this related to children being encouraged to bring their viewpoints and develop new shared understandings. Although the sample size was small, the strong inter-rater reliability and appropriate statistical analysis of the results, resulting in their medium WoE D rating, suggest this study can be considered alongside those included within Hughes and Schlosser's (2014) review.

Conclusions and Recommendations

This review has contributed to the work of Hughes and Schlosser (2014) and the evidence of NGs being an effective intervention that can improve the well-being of children with SEMH needs. Though all studies highlighted improvements in at least one area of a child's social, emotional and mental health, the extent to which their contributions were weighted was in line with the in-depth assessment of their WoE. For example, within the BP developmental model, measuring SEMH factors such as emotional security, large pre-post effect sizes were shown by Grantham and Primrose (2017) and Sloan et al. (2020), and although no effect size was calculated by Lyon (2017) they showed their participants all consistently improved. The latter's participants also all improved within the BP diagnostic profile, likewise Hibbin and Warin (2016) showed 83% of their participants improved in this measure post-test. As this measure details a child's social wellbeing through behaviour these improvements, alongside large pre/post effect sizes found by other studies (Grantham & Primrose, 2017; Sloan et al., 2020), demonstrate NGs positive impact on social development. In addition, Sloan et al.'s (2020) between groups multi-level regression showed large effect sizes for NGs positive impact on both the BP developmental strand and diagnostic profile. Furthermore, Cunningham et al. (2019) reported large and medium effect sizes for NGs impact on two social wellbeing measures post-test, CRPM and TOPSS respectively.

The near perfect WoE D rating for Sloan et al. (2020) should convey the strength of this study, with the main highlights being its considerable sample size as well as the inclusion of a control group allowing for greater reliability of effect size. Their aforementioned large effect sizes found across the BP, as well as across the SDQ which is another strong indicator of the SEMH of young people (Goodman, 1997), indicate NG's effectiveness in improving student's wellbeing. This statement is also supported by large effect sizes found within participants pre and post by Cunningham et al. (2019) and Grantham and Primrose (2017) as they both received medium WoE D ratings. Despite the former's large effect size for NGs on the directly administered CRPM (a child's strength in social situations), their questionable TOPSS results and Grantham and Primrose's (2017) limited significance found from their BP diagnostic scale, suggests that there are limits to the social benefit of NGs. However, this is not supported by the two remaining studies, while both have a low WoE D rating, their findings are in line with Sloan et al. (2020) in regards to a majority across the board BP improvement as well as Grantham and Primrose's (2017) large effect size found for a reduction in "Disengagement".

Though the quality of research varied, the outcomes were not significantly impacted between classic and part time NG models either between studies (Cunningham et al., 2019; Grantham and Primrose, 2017) or within studies (Sloan et al., 2020). Although Grantham and Primrose (2017) and Lyon (2017) contributed to the evidence of NGs in secondary schools, the need for further research in this age group is evident. In addition, the only fair analysis of "classic" vs "adapted" NGs was conducted by Sloan et al. (2020) and they found no significant difference, the lack of varied NG type representation in other studies prevents this review from drawing any wider conclusions on type of NG having an impact on effectiveness.

Although it is evident that randomised control trials are a tall order with the significant existing NG population, further studies following the approach of Sloan et al. (2020) would greatly benefit the field of research. A particular strength being the consideration of confounding variables which have given pause for thought in regard to greater benefits of NGs to those with lower baseline

scores. Another strength is also the inclusion of a control group as this differentiates it from the pre/post studies and challenges the assertion that maturation is the significant factor in change, due to the disparity between the intervention and control groups.

Overall this review presented notable evidence for the effectiveness of NGs on improving wellbeing mainly through the Sloan et al. (2020) study's findings, supported mostly by the two reasonably strong studies conducted by Cunningham et al. (2019) and Grantham and Primrose (2017). However, the pre-post nature of the majority of studies presents the possibility that any improvements seen are as a result of confounding variables as oppose to NGs themselves. Such as the development of student-teacher and/or student-peer relationships or even simply the natural maturation of the children over the time of the study. Therefore, despite adding to the work of Hughes and Schlosser (2014), the implications for practice remain that tentatively NGs can be recommended as an intervention to reduce presenting SEMH needs in children. Through presenting the studies findings after evaluation via the WoE framework, practitioners can benefit from a further explored evidence base to inform their practice and underpin their recommendations for NGs as a suitable intervention.

Research in relation to the secondary review question has also contributed specifically to strategies utilised with success in comparison to mainstream. Namely the increased use of attunement principles that support the core psychology of the NG intervention, attachment theory. This is evident through the building of relationships and secure attachments. However, it is important to note that the only study found by this review that addressed the review question had a very limited sample size (Cubeddu & Mackay, 2017). The recommendations for practice that can be echoed by this review is the promotion of attunement principles for use within NG teaching. However. similar research on a larger scale is recommended to increase the generalisability of the research into strategies within NGs that promote SEMH, especially for the most vulnerable groups. In addition, further qualitative research involving children in NGs that have shown effectiveness in

improving wellbeing, should be conducted to understand the elements of NGs that are consistent with such significantly positive outcomes.

References

Bowlby, J. (1969). *Attachment and Loss*, Vol.1 Attachment. New York: Basic Books.

Boxall, M. (2002). *Nurture Groups: Principles and Practice*. London: Paul Chapman

Cameron, S and Maginn, C, (2011). Living psychology: The 'emotional warmth' dimension of professional childcare. *Educational and Child Psychology*, *28*(3) . 44. https://doi.org/10.1007/978-94-007-6398-2_10

Cannon, Y., Gregory, M., and Waterstone, J. (2013). A solution hiding in plain sight: Special education and better outcomes for students with social, emotional, and behavioral challenges. *The Fordham Urban Law Journal*, *41*(2), 403. https://ir.lawnet.fordham.edu/ulj/vol41/iss2/3/

Child Outcomes Research Consortium (CORC). (2020). *Pupil Attitudes to Self and School.* https://www.corc.uk.net/outcome-experience-measures/pupil-attitudes-to-self-and-school/

Cooke, C, Yeomans, J and Parkes, J (2008) The Oasis: Nurture Group Provision for Key Stage 3 Pupils, *Emotional and Behavioural Difficulties*, *13*(4). 291-303. https://doi.org/10.1080/13632750802442219

Croft, S., C. Stride, B. Maughan, and R. Rowe. (2015). Validity of the Strengths and Difficulties Questionnaire in Preschool-Aged Children. *Pediatrics*. *135*(5): e1210–e1219. https://doi.org/10.1542/peds.2014-2920

Cubeddu, D. and MacKay, T. (2017) The attunement principles: a comparison of nurture group and mainstream settings*. Emotional and Behavioural Difficulties*. *22*(3), 261-274. https://doi.org/10.1080/13632752.2017.1331985

Cunningham, L., Hartwell, B. and Kreppner, J. (2019) Exploring the impact of Nurture Groups on children's social skills: a mixed-methods approach*. Educational Psychology in Practice*. *35*(4). 368-383. https://doi.org/10.1080/02667363.2019.1615868

Department for Education (DfE). (2018). Mental health and behaviour in schools. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/755135/Mental_health_and_behaviour_in_schools__.pdf

Department for Education (DfE) and Department of Health (DoH). (2014). Special educational needs and disability code of practice: 0 to 25 years. London: DfE/DoH. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/398815/SEND_Code_of_Practice_January_2015.pdf

Dodge, K.A., McClaskey, C.L. and Feldman, E. (1985). Situational approach to the assessment of social competence in children. *Journal of Consulting and Clinical Psychology*, *53*(3), 344-353. https://doi.org/10.1037//0022-006x.53.3.344

Downs SH and Black N. (1998). The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *Journal of Epidemiological Community Health*. *52*(6). 377-84. https://doi.org/10.1136/jech.52.6.377.

Early Intervention Foundation. (2017). GUIDEBOOK: Nurture Groups. https://guidebook.eif.org.uk/programme/nurture-groups

Glen, S. (2016)."Hedges' g: Definition, Formula. StatisticsHowTo.com: Elementary Statistics for the rest of us! https://www.statisticshowto.com/hedges-g/

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A Research Note. *Journal of Child Psychology and Psychiatry*, *38*(5), 581–586. https://doi.org/10.1037/t00540-000

Gough, D. (2007). Weight of evidence: a framework for the appraisal of the quality and relevance of evidence. *Research Papers in Education*, *22*(2), 37–41. https://doi.org/10.1080/02671520701296189

Grantham, R and Primrose, F. (2017) Investigating the fidelity and effectiveness of Nurture Groups in the secondary school context. *Emotional and Behavioural Difficulties. 22*(3), 219-236. https://doi.org/10.1080/13632752.2017.1331986

Hibbin, R and Warin, J. (2016). Nurture groups in practice: children; classes; schools: Final report of Comparative study of nurture groups and alternative provisions for children with social, emotional and behavioural difficulties. *Nurture Group Network.* https://doi.org/10.13140/RG.2.1.3477.1609

Hughes, N.K. and Schlösser, A. (2014) The effectiveness of nurture groups: a systematic review, *Emotional and Behavioural Difficulties, 19*(4), 386-409, https://doi.org/10.1080/13632752.2014.883729

Jacobson, N. S., and Truax, P. (1991). Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *Journal of consulting and clinical psychology, 59*(1), 12. https://doi.org/10.1037//0022-006x.59.1.12

Lyon, L. (2017). A pilot study of the effectiveness of a nurture group in a secondary special school. *International Journal of Nurture in Education*, *3*(1), 6-17. https://www.nurtureuk.org/sites/default/files/lyon.pdf

Maslow, A. H. (1970). *Motivation and Personality.* 2nd ed. London: Harper and Row.

Nangle, D, Ellis, J. and Hansen, D. (1994). Clinical Utility of the Taxonomy of Problematic Social Situations for Children (TOPS): Temporal Stability and Convergence with Peer Measures. *Education and Treatment of Children*. *17*. 129-138. https://www.researchgate.net/profile/David-Hansen-11/publication/234619464_Clinical_Utility_of_the_Taxonomy_of_Problematic_Social_Situations_for_Children_TOPS_Temporal_Stability_and_Convergence_with_Peer_Measures/links/53fb7e6f0cf22f21c2f33122/Clinical-Utility-of-the-Taxonomy-of-Problematic-Social-

Situations-for-Children-TOPS-Temporal-Stability-and-Convergence-with-Peer-Measures.pdf

NHS Digital. (2020). Mental Health of Children and Young People in England: Wave 1 follow up to the 2017 survey. https://files.digital.nhs.uk/AF/AECD6B/mhcyp_2020_rep_v2.pdf

Nurture UK. (2019). *50 years of nurture*. https://www.nurtureuk.org/50-years-nurture

Ofsted. (2011). *Supporting Children with Challenging Behaviour Through a Nurture Group Approach.* Manchester: Crown Copyright. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/413529/Supporting_children_with_challenging_behaviour_through_a_nurture_group_approach.pdf

Petticrew, M., and Roberts, H. (2003). Evidence, hierarchies, and typologies: horses for courses. *Journal of Epidemiology Community Health*, *57*(7), 527-529. https://doi.org/10.1136/jech.57.7.527

RAND. (2020). Pupil Attitudes to Self and School (PASS). RAND Education Assessment Finder. https://www.rand.org/education-and-labor/projects/assessments/tool/2000/pupil-attitudes-to-self-and-school-pass.html

Seth-Smith, F., Netali L., Richard P., Fonagy, P. and Jaffey, D. (2010), 'Do nurture groups improve the social, emotional and behavioural functioning of at risk children?' *Educational and Child Psychology*, *27*(1), 21-34. http://growinggreatschoolsworldwide.com/wp-content/uploads/2018/02/ECP27_1-Seth-Smith-et-al.pdf

Sloan, S, Winter, K, Connolly, P, and Gildea, A. (2020). The effectiveness of Nurture Groups in improving outcomes for young children with social, emotional and behavioural difficulties in primary schools: An evaluation of Nurture Group provision in Northern Ireland. *Children and Youth Services Review*, *108*, 104619. https://doi.org/10.1016/j.childyouth.2019.104619

Appendix A – Mapping the field

No	Author	Country	Participants	Study Design	Nurture Group Method	Measures	Primary Outcomes
1	Cubeddu and Mackay (2017)	Scotland (UK)	N=5 One NG teacher and four Mainstream teachers. All teachers were female. Teachers all taught classes of 5-6 year olds.	Observational Study – Event Sampling	"Classic" Morning weekday NG.	Event Sampling of the six attunement principles: Being attentive Encouraging initiatives Receiving initiatives Developing attuned interactions Guiding Deepening discussion.	The NG teacher was observed to utilise the 6 principles more frequently and evenly than mainstream counterparts. Mainstream teachers most often utilised "guiding", this may be representative of typical in class support or "hand holding".
2	Cunningham et al. (2019)	England (UK)	N=16 aged between 6 and 9.75 years old. Majority also receiving additional phonics, handwriting and numeracy support. 15 white British, 1 Asian. Two receiving occupational therapy. Three children were from military families	Mixed methods approach	"Adapted" (part time). Post measures taken after 15 weeks of NG intervention	Child Role Play Measure (CRPM) (a=0.72) – Administered by Author, child asked what they would do in 15 different social situations, responses scored. In this study time 1 a=0.47 and time 2 a=0.79. Taxonomy of Problematic Social Situations (TOPSS) (a=0.79) – completed by class teachers, 44 items related to the child's performance in social situations. In this study time	Significant change (p=0.002)in CRPM following NG with a large effect size (d=0.97). However low reliability (a=0.47) at time 1 raises concerns. Only approaching significance (p=0.055) in TOPSS with a moderate effect size (d=0.52). However RCI shows 56% positive improvement, 25% no improvement and 19% regression of social situation performance.

No	Author	Country	Participants	Study Design	Nurture Group Method	Measures	Primary Outcomes
						1 a=0.97 and time 2 a=0.98. Qualitative interviews – semi-structured interviews conducted by the author to the children.	
3	Grantham and Primrose (2017)	Scotland (UK)	N=24. Secondary aged pupils who attended a NG in the 2014-15 academic year. Participants came from 5 different schools.	Mixed Methods approach	"Adapted" "Secondary Nurture Bases"	<u>Qualitative:</u> Six structured interview questions Six structed questionnaires <u>Quantitative</u> Boxall Profile for Young People (BPYP) – split into 10 developmental strand scores and 10 diagnostic profile scores.	Effect sizes (η^2 Large and statistically significant for an increase in 8 of the 10 developmental strands. However, the same can only be said for one (disengaged) in terms of significance and effect size of the decrease. Developmental areas of specific increase were noted as the social aspects of children's Interaction within the classroom and making friends. Also, a developmental strand indicative of resilience was found to be significant. However, while decreases in social and

No	Author	Country	Participants	Study Design	Nurture Group Method	Measures	Primary Outcomes
							emotional diagnostic scores were present, they were not statistically significant.
4	Hibbin and Warin, (2016)	England (UK)	N = 14, Across 7 education settings, two case studies from each. Settings were separated by 5 being primary schools and 2 alternative provisions. 1 AP had recently stopped its NG provision, the other was residential for children with SEMH needs. Ages ranged from 6 – 11. All had differing lengths of time in NG provision	Mixed Methods Approach	Varied between settings	<u>Qualitative</u> Interviews with Head Teachers, Staff, child and parent/carers Focus groups involving NG and mainstream staff. Classroom observations **<u>Quantitative</u>** N=12 - Boxall Profiles across various timepoints for the target children. Reported only the Diagnostic Indicator scores. N=2 - Qualifications and Curriculum Authority's Assessment of Learning, Conduct and Emotional Behaviour (QCA CEB). Recorded across at least 3 time points.	Trends appeared to show improvement in social and emotional functioning overtime within the settings. Success was correlated with high levels of communicatio n between mainstream and NG teaching staff as well as minimal contrast between context of NG vs mainstream.
5	Lyon (2017)	England (UK)	N=4, three aged between	Mixed Methods Approach	"Classic" – progress of students	<u>Qualitative:</u>	Three out of the four pupils showed

No	Author	Country	Participants	Study Design	Nurture Group Method	Measures	Primary Outcomes
			11-12 and one aged 12-13, all male.		monitored over an academic year.	Semi-structed interviews with pupils Questionnaires relating to the perceived effectiveness of NG were administered to school staff, NG facilitators and parents/ carers. <u>Quantitative</u> Boxall Profile. Completed across three time points. Profile and Pupil Attitude to Self and School (PASS). Completed across three time points Structured observations – interval recording (every 5 minutes for 45 minutes). Completed across four time points in both mainstream and NG settings.	sustained improvement in attention in mainstream based on observations. All four pupils made progress in terms of social and emotional development over the course of the year. Three out of four of the students showed progress in their PASS score – Self-esteem and academic self-concept.
6	Sloan et al. (2020)	Northern Ireland (UK)	(Signature Project funding - schools with a higher than average population	Non-randomised Control Trial	Mixed between classic (full time) and adapted (part time).	<u>Quantitative</u> Boxall Profile – Developmental (a=0.931) and Diagnostic (a=0.919)	Children in the intervention group showed mean significant improvements across both the BP and

No	Author	Country	Participants	Study Design	Nurture Group Method	Measures	Primary Outcomes
			of children who get FSM, have below average attendance and attainment. As well as that, at ages 5–6 and 10–11 years, tests show above average levels of SEN children) N=384, 232 across 20 "Signature Project" Nurture Groups, 66 from 10 "established Nurture Groups" and 88 from 14 "Signature "Project equivalent schools with no NG provision.			were measured pre and post intervention. Strengths and Difficulties Questionnaire (SDQ) – Main total difficulties (a=0.798) and prosocial behaviour (a=0.818) scores were measured pre and post intervention. Bespoke self-report relating to Enjoyment of school, child self-report. (a=0.799). KS1 and KS2 assessment tests on Maths and ICT.	SDQ. Compared low evidence of control group change. Only small, insignificant, changes were noted in relation to attitude to school and academic assessments measured. Nurture Groups effects were large for the developmental strand (g = 1.352) and the diagnostic profile (g = -0.904) of BP and for the total difficulties score (g = -1.303) and prosocial scores (g = 0.926) of the SDQ Confounding variables included: Larger school size having a negative correlation with peer problems for the children. Those with the lowest baseline scores saw the most significant positive

No	Author	Country	Participants	Study Design	Nurture Group Method	Measures	Primary Outcomes
							effects of the NG intervention. .

Appendix B – Excluded studies

Study	Exclusion Criteria
Allison, J. and Shirley, C. (2014). Keeping Our Difficult Kids in School: The Impact of the Use of the 'Boxall Profile' on the Transition and Integration of Behaviourally-Disordered Students in Primary Schools. https://files.eric.ed.gov/fulltext/EJ1240596.pdf	2,4,5
Bailey, S. (2014). Exploring ADHD: An ethnography of disorder in early childhood. http://dx.doi.org/10.4324/9780203119723	2,4,5
Barn, J. (2015). 60 Mindful Minutes: developing mindful behaviour in the nurture group. *Educational Psychology in Practice. 31*(3): 330-331. https://doi.org/10.1080/02667363.2015.1052232	4,5
Birch, E. (2016). 'You do what you need for your children, don't you?': An exploration of the current range of practice and priorities of nurture group staff in a local authority. *Educational and Child Psychology. 33*(4), 40-49.	4,5
Blanco-Bayo, A. (2020). "It doesn't matter because I love you". A case study examining the interpretation of Behaviour Classification Tables and Positive Behaviour Support models. *Emotional and Behavioural Difficulties 25*(2), 155-168. https://doi.org/10.1080/13632752.2020.1738698	2,4,5
Carroll, C. and Hurry, J. (2018). Supporting pupils in school with social, emotional and mental health needs: a scoping review of the literature. *Emotional and Behavioural Difficulties 23*(3): 310-325. https://doi.org/10.1080/13632752.2018.1452590	2,4,5
Cefai, C, Arlove, A, Duca, M, Galea, N, Muscat, M. and Cavioni, V. (2018) RESCUR Surfing the Waves: an evaluation of a resilience programme in the early years, *Pastoral Care in Education*, *36*(3), 189-204, https://doi.org/10.1080/02643944.2018.1479224	2,4,5

Cefai, C. and Pizzuto, S.A.S. (2017). Listening to the voices of young children in a nurture class. *Emotional and Behavioural Difficulties 22*(3): 248-260. https://doi.org/10.1080/13632752.2017.1331987	4,5 (Full text read)
Cheney, G, SchoLosser, A. and Nash, P. (2014). Targeted group-based interventions in schools to promote emotional well-being: A systematic review. *Clinical Child Psychology and Psychiatry 19*(3): 412-438. https://doi.org/10.1177%2F1359104513489565	2,4,5
Coleman, M. (2020). Leading the change to establish a whole-school nurturing culture. *Emotional and Behavioural Difficulties 25*(1): 68-79. https://doi.org/10.1080/13632752.2019.1682244	4,5
Coleman, M. and Cooper P. (2017). Nurture and nurture groups. *Emotional and Behavioural Difficulties 22*(3): 185-187. https://doi.org/10.1080/13632752.2017.1335114	4,5
Cooper, P. (2017). Building Social-Emotional Resilience in Schools. *Life in Schools and Classrooms*. 489-506. https://link.springer.com/chapter/10.1007/978-981-10-3654-5_30	4,5
Daniels, H. (2014). Editorial, Emotional & Behavioural Difficulties. 19(4), 341-342.	4,5
Davison, P. and Duffy, J. (2017). A model for personal and professional support for nurture group staff: to what extent can group process consultation be used as a resource to meet the challenges of running a nurture group? *Educational Psychology in Practice 33*(4): 387-405. https://doi.org/10.1080/02667363.2017.1336704	4,5
Delafield-Butt, J. (2018). The emotional and embodied nature of human understanding: Sharing narratives of meaning. *The Child's Curriculum*, 59-84	2,4,5

Delafield-Butt, J and Dunlop, AW,. (2018). *The child's curriculum: Working with the natural values of young children*. New York, NY, US: Oxford University Press; US.	2,4,5
Delafield-Butt, J. T. and Adie J. (2016). The Embodied Narrative Nature of Learning: Nurture in School. *Mind Brain and Education 10*(2): 117-131. https://dx.doi.org/10.1111/mbe.12120	4,5
Duan, W, Bu, H and Chen, Z. (2020). COVID-19-related stigma profiles and risk factors among people who are at high risk of contagion. *Soc Sci Med*. https://doi.org/10.1016/j.113425.	2,4,5
Ennals, P, Fortune, T, Williams, A and D'Cruz, K. (2016) Shifting occupational identity: doing, being, becoming and belonging in the academy, *Higher Education Research & Development*, *35*(3), 433-446, https://doi.org/10.1080/07294360.2015.1107884	2,4,5
Griffiths, R., Stenner, R. and Hicks, U. (2014). Hearing the unheard: Children's constructions of their Nurture Group experiences. *Educational and Child Psychology. 31*. 124-136.	4,5
Hibbin, R. and J. Warin (2020). A language focused approach to supporting children with social, emotional and behavioural difficulties (SEBD). *Education 3-13 48*(3): 316-331. http://dx.doi.org/10.1080/03004279.2019.1664410	4,5
Hughes, N. K. and Schlosser, A. (2014). The effectiveness of nurture groups: A systematic review. *Emotional & Behavioural Difficulties. 19*(4), 386-409. http://dx.doi.org/10.1080/13632752.2014.883729	4,5
Kirkbride, R. (2014). "They Were a Little Family": An Exploratory Study of Parental Involvement in Nurture Groups--From a Practitioner and Parent Perspective. *British Journal of Special Education*. *41*(1): 82-104. https://doi.org/10.1111/1467-8578.12047	4,5

Lavoie, C, Couture, C., Begin, JY and Masse, L. (2017). The differentiated impact of Kangaroo Class programmes in Quebec primary schools: examining behavioural improvements in relation to student characteristics. *Emotional and Behavioural Difficulties 22*(3): 275-289. https://doi.org/10.1080/13632752.2017.1335118	2,4,5
March, S. and Kearney, M. (2017). A psychological service contribution to nurture: Glasgow's nurturing city. *Emotional and Behavioural Difficulties 22*(3): 237-247. http://dx.doi.org/10.1080/13632752.2017.1331972	4,5
Pace, M. (2020). BUILDING RESILIENCE SKILLS THROUGH JOURNAL WRITING AND DRAWING.	2,4,5
Rawdin, C. (2019). Critical realism with a 'small-c': using domain theory to conceptualize therapeutic education. *Journal of Critical Realism 18*(2): 123-141. https://doi.org/10.1080/14767430.2019.1599558	2,4,5
Sackville-Ford, M. (2019). Behaviours that challenge at Forest School.	4,5
Sackville-Ford, M and Davenport, H. (2019). *Critical issues in Forest Schools.* Thousand Oaks, CA: Sage Publications Ltd.	4,5
Salisbury, S. (2018). Using attachment enhancing activities based on the principles of Theraplay (R) to improve adult-child relationships and reduce a child's 'overall stress' as measured by the Strength and Difficulties Questionnaire (SDQ). *Emotional and Behavioural Difficulties 23(*4): 424-440. https://doi.org/10.1080/13632752.2018.1497000	2,4,5
Segev, R. and Shoham, A. (2016). The social and dual identity role of joint gift-giving among adolescents. Young Consumers 17(1): 3-17. https://www.emerald.com/insight/content/doi/10.1108/YC-07-2015-00542/full/html	4,5

Sossick, M., Gumbrell, D. and Allen, P. (2019). The Impact of a Coaching Project on the Resilience of Newly Qualified Teachers. *Teacher Education Advancement Network Journal. 11*(1). 15-24. https://ojs.cumbria.ac.uk/index.php/TEAN/article/view/502/622	4,5
Stone, K., Burgess, C, Daniel, B, Smith, J. and Stephen, C. (2017). Nurture corners in preschool settings: involving and nurturing children and parents. *Emotional and Behavioural Difficulties 22*(4): 383-396. https://doi.org/10.1080/13632752.2017.1309791	4,5 (Full text read)
Sutterlin, C. (2019). Culture by Nature. Familial roots of ambivalent human social behavior and its cultural extensions in large-scale societies. A contribution of Human and Cultural Ethology. *Anthropologischer Anzeiger, 76*(3), 195-210. https://doi.org/10.1127/anthranz/2019/0883	2,4,5
Vincent, K. (2017). It's Small Steps, but That Leads to Bigger Changes: Evaluation of a Nurture Group Intervention. *Emotional and Behavioural Difficulties, 22*(4), 303-316. https://doi.org/10.1080/13632752.2017.1290882	4,5 (Full text read)
Warin, J. (2017). Creating a whole school ethos of care. *Emotional and Behavioural Difficulties 22*(3): 188-199. https://doi.org/10.1080/13632752.2017.1331971	4,5 (Full text read)
Warin, J. (2018). Men and Care: Diversity and Equity.*Early Childhood Education and Care.* 79-98. https://doi.org/10.1007/978-3-319-89539-0_5	2,4,5

Appendix C – WoE A

Due to the nature of the studies included being quasi-experimental the Downs and Black Checklist (1998) was utilised. In line with Hughes and Schlosser (2014) several of the items were deleted due to their incompatibility with the studies, the total potential score was adjusted accordingly. Reasons for deleting the items were; participants inability to have meta-cognition of their participation (11,12,14 and 24) and through conducting the study the researchers were unable to comply with the items (15,16,19 and 23). Furthermore, item 13 was ratified to be specific to the location of the Nurture Group intervention. Finally, items 9 and 26 were reduced to “any time point” as oppose to “follow up”. This was because all studies took the form of pre and post measurement. Figure C 1 shows an example of the completed checklist for one of the studies. The WoE was calculated by taking the total score out of 24 and dividing it by 8 to calculate a score where the maximum would be 3. Table C 1 shows the WoE A rating of each study.

Figure C 1:

Cubeddu and Mackay (2017)

Reporting

1. *Is the hypothesis/aim/objective of the study clearly described?*

yes	1

2. *Are the main outcomes to be measured clearly described in the Introduction or Methods section?*

If the main outcomes are first mentioned in the Results section, the question should be answered no.

yes	1

3. *Are the characteristics of the patients included in the study clearly described ?*

In cohort studies and trials, inclusion and/or exclusion criteria should be given. In case-control studies, a case-definition and the source for controls should be given.

yes	1

4. *Are the interventions of interest clearly described?*

Treatments and placebo (where relevant) that are to be compared should be clearly described.

yes	1

5. *Are the distributions of principal confounders in each group of subjects to be compared clearly described?*

A list of principal confounders is provided.

no	0

6. *Are the main findings of the study clearly described?*

Simple outcome data (including denominators and numerators) should be reported for all major findings so that the reader can check the major analyses and conclusions. (This question does not cover statistical tests which are considered below).

yes	1

7. *Does the study provide estimates of the random variability in the data for the main outcomes?* In non normally distributed data the inter-quartile range of results should be reported. In normally distributed data the standard error, standard deviation or confidence intervals should be reported. If the distribution of the data is not described, it must be assumed that the estimates used were appropriate and the question should be answered yes.

yes	1

8. *Have all important adverse events that may be a consequence of the intervention been reported?* This should be answered yes if the study demonstrates that there was a comprehensive attempt to measure adverse events. (A list of possible adverse events is provided).

no	0

9. *Have the characteristics of patients lost at any time point been described?*

This should be answered yes where there were no losses to follow-up or where losses to follow-up were so small that findings would be unaffected by their inclusion. This should be answered no where a study does not report the number of patients lost to follow-up.

yes	1

10. *Have actual probability values been reported(e.g.0.035 rather than <0.05) for the main outcomes except where the probability value is less than 0.001?*

yes	1

External validity

All the following criteria attempt to address the representativeness of the findings of the study and whether they may be generalised to the population from which the study subjects were derived.

11. ~~*Were the subjects asked to participate in the study representative of the entire population from which they were recruited?*~~

~~The study must identify the source population for patients and describe how the patients were selected. Patients would be representative if they comprised the entire source population, an unselected sample of consecutive patients, or a random sample. Random sampling is only feasible where a list of all members of the relevant population exists. Where a study does not report the proportion of the source population from which the patients are derived, the question should be answered as unable to determine.~~

~~yes~~	~~1~~
~~no~~	~~0~~
~~unable to determine~~	~~0~~

12. ~~*Were those subjects who were prepared to participate representative of the entire population from which they were recruited?*~~

~~The proportion of those asked who agreed should be stated. Validation that the sample was representative would include demonstrating that the distribution of the main confounding factors was the same in the study sample and the source population.~~

~~yes~~	~~1~~
~~no~~	~~0~~
~~unable to determine~~	~~0~~

13. *Did the nurture group take place in a school environment?* For the question to be answered yes the study should demonstrate that the intervention was representative of that in use in the source

population. The question should be answered no if, for example, the intervention was undertaken in a specialist centre unrepresentative of the hospitals most of the source population would attend.

yes	1

Internal validity - bias

14. ~~*Was an attempt made to blind study subjects to the intervention they have received ?*~~
~~For studies where the patients would have no way of knowing which intervention they received, this should be answered yes.~~

~~yes~~	~~1~~
~~no~~	~~0~~
~~unable to determine~~	~~0~~

15. ~~*Was an attempt made to blind those measuring the main outcomes of the intervention?*~~

~~yes~~	~~1~~
~~no~~	~~0~~
~~unable to determine~~	~~0~~

16. ~~*If any of the results of the study were based on "data dredging", was this made clear?* Any analyses that had not been planned at the outset of the study should be clearly indicated. If no retrospective unplanned subgroup analyses were reported, then answer yes.~~

~~yes~~	~~1~~
~~no~~	~~0~~
~~unable to determine~~	~~0~~

17. *In trials and cohort studies, do the analyses adjust for different lengths of follow-up of patients, or in case-control studies, is the time period between the intervention and outcome the same for cases and controls ?* Where follow-up was the same for all study patients the answer should yes. If different lengths of follow-up were adjusted for by, for example, survival analysis the answer should be yes. Studies where differences in follow-up are ignored should be answered no.

unable to determine	0

18. *Were the statistical tests used to assess the main outcomes appropriate?*
The statistical techniques used must be appropriate to the data. For example nonparametric methods should be used for small sample sizes. Where little statistical analysis has been undertaken but where there is no evidence of bias, the question should be answered yes. If the distribution of the data (normal or not) is not described it must be assumed that the estimates used were appropriate and the question should be answered yes.

yes	1

19. ~~*Was compliance with the intervention/s reliable?*~~
~~Where there was non compliance with the allocated treatment or where there was contamination of one group, the question should be answered no. For studies where the effect of any misclassification was likely to bias any association to the null, the question should be answered yes.~~

~~yes~~	~~1~~
~~no~~	~~0~~

~~unable to determine~~	~~0~~

20. *Were the main outcome measures used accurate (valid and reliable)?*

For studies where the outcome measures are clearly described, the question should be answered yes. For studies which refer to other work or that demonstrates the outcome measures are accurate, the question should be answered as yes.

yes	1

Internal validity - confounding (selection bias)

21. *Were the patients in different intervention groups (trials and cohort studies) or were the cases and controls (case-control studies) recruited from the same population?* For example, patients for all comparison groups should be selected from the same hospital. The question should be answered unable to determine for cohort and casecontrol studies where there is no information concerning the source of patients included in the study.

yes	1
unable to determine	0

22. *Were study subjects in different intervention groups (trials and cohort studies) or were the cases and controls (case-control studies) recruited over the same period of time?* For a study which does not specify the time period over which patients were recruited, the question should be answered as unable to determine.

yes	1

~~23. *Were study subjects randomised to intervention groups?*~~

~~Studies which state that subjects were randomised should be answered yes except where method of randomisation would not ensure random allocation. For example alternate allocation would score no because it is predictable.~~

~~yes~~	~~1~~
~~no~~	~~0~~
~~unable to determine~~	~~0~~

~~24. *Was the randomised intervention assignment concealed from both patients and health care staff until recruitment was complete and irrevocable?*~~

~~All non-randomised studies should be answered no. If assignment was concealed from patients but not from staff, it should be answered no.~~

~~yes~~	~~1~~
~~no~~	~~0~~
~~unable to determine~~	~~0~~

25. *Was there adequate adjustment for confounding in the analyses from which the main findings were drawn?*

This question should be answered no for trials if: the main conclusions of the study were based on analyses of treatment rather than intention to treat; the distribution of known confounders in the different treatment groups was not described; or the distribution of known confounders differed between the treatment groups but was not taken into account in the analyses. In nonrandomised

studies if the effect of the main confounders was not investigated or confounding was demonstrated but no adjustment was made in the final analyses the question should be answered as no.

no	0

26. *Were losses of patients to at any time point taken into account?*

If the numbers of patients lost to follow-up are not reported, the question should be answered as unable to determine. If the proportion lost to follow-up was too small to affect the main findings, the question should be answered yes.

yes	1

Power

27. *Did the study have sufficient power to detect a clinically important effect where the probability value for a difference being due to chance is less than 5%?*

Sample sizes have been calculated to detect a difference of x% and y%.

	Size of *smallest* intervention group	
A	$<n_1$	0

Total = 14/24

WoE A= 1.75/3 (Total / 8)

Table C 1 – WoE A scores:

Study	WoE A
Cubeddu and Mackay (2017)	1.8
Cunningham et al. (2019)	2.1
Grantham and Primrose (2017)	1.9
Hibbin and Warin, (2016)	0.9
Lyon (2017)	1.1
Sloan et al. (2020)	3

Appendix D – WoE B

WoE B is adjudged specifically to the individual review and relates to the appropriateness of the article's contents to the review question (Gough, 2007). As effectiveness is the key measure it was logical to utilise Petticrew and Roberts (2003) hierarchies, the studies should therefore follow their below system of weighting based on their design. Their hierarchies of design have been adapted into the WoE B rating scale (Table D 1) and then applied to the six reviewed studies (Table D 2).

Table D 1- WoE B rating criteria

WoE B Rating and Qualitative Descriptor	Criteria
1 (Low)	Non-experimental evaluation, qualitative research, case control or survey
2 (Medium)	Quasi-experimental studies, cohort studies, single case experimental designs
3 (High)	Randomised Control Trials

Table D 2 - WoE B scores:

Study	Overall WoE B
Cubeddu and Mackay (2017)	2
Cunningham et al. (2019)	2
Grantham and Primrose (2017)	2
Hibbin and Warin, (2016)	2
Lyon (2017)	2
Sloan et al. (2020)	3

Appendix E – WoE C

WoE C is utilised to evaluate the study's level of relevance to the review question and its topic (Gough, 2007). The criteria used and accompanying rating scales and rational can be found in Table E 1. WoE C rating for the six reviewed studies are shown in Table E 2.

Table E 1 – Rating scale rationale:

Criteria	WoE C Rating	Descriptor	Rationale
Objectivity of the study	1	Objectivity not discussed, or reason to believe the author has a vested interest	Educational Psychologists have a duty to offer independent and impartial advice, this extends to the interventions they recommended. It is vital that evidence utilised to recommend interventions is free of bias to ensure merit and efficacy are championed as key impartial indicators.
	2	Objectivity discussed but potential for vested interest	
	3	Objectivity discussed and no reason to believe the author has a vested interest.	
Outcome measures reliability and validity	1	Outcome measure does not report/ or reports low, reliability or validity	The reliability and validity of outcome measures are vital in their ability to fairly access the effectiveness of an intervention on the targeted difficulty.
	2	Outcome measure reports moderate reliability and validity	
	3	Outcome measure reports high reliability and validity	

Participant confounding variables considered	1	Participant confounding variables are not reported or recorded.	When evaluating an intervention, such as NGs, it is vital that the demographic data of the participants, such ethnicity, socio-economic class and underlying SEN is identified to limit the alternative hypothesis for improvement that could be concluded. Therefore, more accurately accessing the effectiveness of the intervention itself.
	2	Participant demographic data is recorded and discussed.	
	3	Participant demographic data is included within the analysis.	
Social, Emotional and Mental Health Needs targeted thoroughly by outcome measure	1	Only one element of Social, Emotional and Mental Health Needs targeted	While showing improvement in any one element of a child's Social, Emotional or Mental Health Needs development demonstrates the strength of an intervention, the inclusion of outcome measures that make reference to the subsections offer greater detail in the evidence base of the intervention.
	2	More than one element of Social Emotional and Mental Health Needs targeted	
	3	All three Social, Emotional and Mental Health Needs aspects targeted through outcome measure.	

Table E 2 - WoE C scores

Study	Objectivity	Outcome Measures	Participants	SEMH difficulties	WoE C
Cubeddu and Mackay (2017)	3	1	1	1	1.5
Cunningham et al. (2019)	3	3	2	1	2.3
Grantham and Primrose (2017)	3	1	1	3	2
Hibbin and Warin, (2016)	1	1	1	1	1
Lyon (2017)	1	1	1	3	1.5
Sloan et al. (2020)	2	3	3	3	2.8