(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

Rationale and relevance to educational psychology practice

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

	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

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

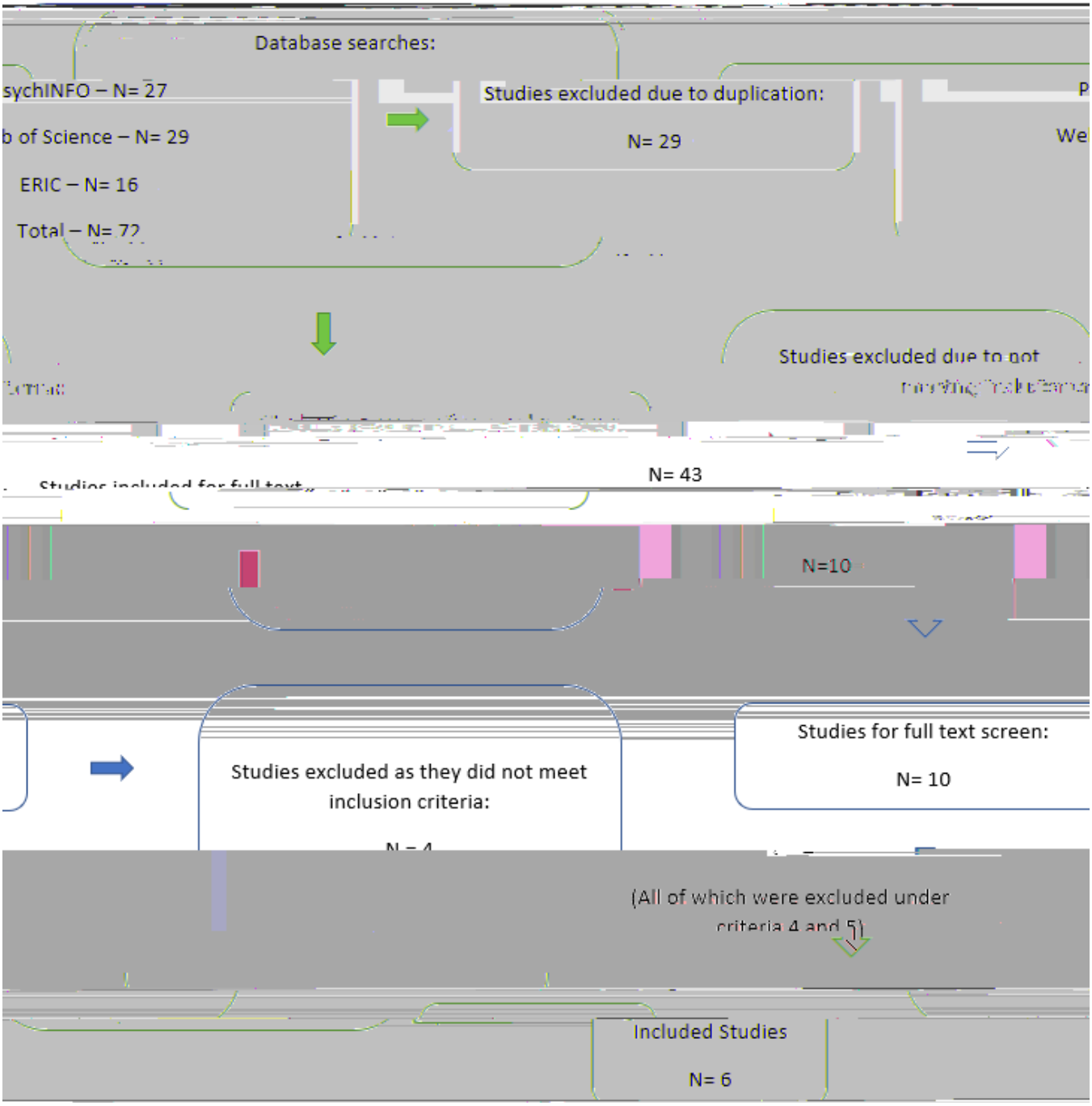
5 Lyon, L. (2017). A pilot study of the effectiveness of a nurture group in a secondary special school*. International Journal of Nurture in Education*, *3*(1), 6-17. https://www.nurtureuk.org/sites/default/files/lyon.pdf

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

differentiate

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.

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,

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

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)					

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,

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.

The near perfect WoE D rating for Sloan et al. (2020) should convey the strength of this study, with th

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

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

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.

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.				

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 soc

No	Author	Country	Participants	Study Design

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
							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	

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):	

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. Wa	

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	

Appendix C

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?*~~

~~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

no	0

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

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