Case Study 1: An Evidence-Based Practice Review Report

Theme: School/setting Based Interventions for Learning.

How effective is Positive Behaviour Support at improving academic achievement in school aged children attending mainstream schools?

Summary

Positive Behaviour Support (PBS) also known as Positive Behavioural Interventions and Supports (PBIS) (Sugai & Horner, 2006) is 'the application of positive behavioural interventions and systems' aimed at achieving socially important behavioural change (Sugai et al., 2000, p.4). Through the application of a three-tiered framework of supports, PBS is a data driven, decision based strategy that applies behavioural and social learning to the whole school context (Bradshaw, Waasdorp, & Leaf, 2015). It aims to proactively prevent challenging and disruptive behaviours in schools (Bradshaw et al., 2015). The primary tier adopts school-wide components. The secondary and tertiary tiers target groups and individuals, respectively, who may require more intensive support than offered by the previous tier (Simonsen, Sugai, & Negron, 2008). PBS has been linked to decreases in inappropriate behaviours, discipline referrals and exclusions (Simonsen et al., 2008). It has also been found to reduce noise levels in schools (Kartub Taylor-Greene, March & Horner, 2000). This review examines the effectiveness of PBS at improving academic achievement in children attending mainstream schools. It identifies mixed findings in effectiveness,

suggesting that its impact on academic achievement is dependent on other variables, such as the quality of teaching, learning support and fidelity.

Introduction

Positive Behaviour Support

As a whole school strategy, PBS aims to build a positive school culture, and to provide support in the learning environment for pupils and teachers (Kelm, McIntosh, & Cooley, 2014). It integrates four critical elements: pupil outcomes, research validated practices that support pupil behaviour, data collection that supports decision making, and school systems that promote each of these (Sugai & Horner, 2002). PBS has been developed on the premise that children require appropriate modelling, monitoring, opportunities to succeed and meaningful feedback that helps to guide behaviour (Lewis & Sugai, 1999). It is a 'noncurricular prevention strategy' offering schools a three-tiered model (Bradshaw et al., 2015, p546). Tier 1 adopts school wide components, while Tiers 2 and 3 complement the whole school approach by providing targeted support and interventions to groups and individuals (Sugai & Horner, 2009).

Beaman and Wheldall (2000) emphasise that the deployment of contingent teacher praise/approval and reprimand/disapproval has 'unequivocally demonstrated' increases in academic and socially appropriate behaviours (p.431). PBS is an approach which focuses on praise/approval. It is a 'compilation of effective practices, interventions, and systems change strategies' (Sugai et al., 2004, p.10). The intention is to remove rewards 'that inadvertently maintain problem behaviour' (Sugai et al., 2000, p. 16), by replacing them with clearly defined behavioural expectations.

Implementation requires schools to identify meaningful outcomes, to establish whole school systems, implement context-appropriate practices, that are evidence based and gather data for decision-making (Simonsen et al., 2008). Outcomes should include goals that schools aim to achieve. The school system must include a team that guides implementation, including a PBS “coach”, who develops and reinforces an action plan (Simonsen et al., 2008). Typically, the coach is experienced at using the functional behavioural assessment method (Bradshaw, Mitchell & Leaf, 2010). Staff training is recommended (Simonsen et al., 2008) and effectiveness is considered to be contingent on staff commitment (Gagnon, Rockwell & Scott., 2008).

Prior to developing an action plan, teachers are encouraged to assess the existing school environment to identify common problem behaviours (Gagnon et al., 2008). Exclusion records or detention data can be evaluated. These records offer quantitative data (frequency and severity of discipline) and qualitative data (descriptions of the behaviour and antecedents). Data analysis is used to identify students who may benefit from higher tier PBS and to pinpoint behaviours that are targeted at the whole school level.

A teaching matrix of specific behavioural expectations is developed through action planning (Kelm et al., 2014), to seek approaches that reinforce positive behaviours (Gagnon, et al., 2008). Students are given specific praise and acknowledgment for demonstrating prosocial behaviour (Kelm et al., 2014). For instance, Chu (2015) describes a token economy where pupils can earn tokens for demonstrating target behaviours. A reinforcer, such as a prize or

reward is offered, after an agreed quantity of tokens. Consequences are given to address challenging behaviours. These are addressed instructionally and consistently, rather than punitively (Kelm et al., 2014). Data collection is an ongoing process, with fidelity being a critical measure that is monitored. There are a number of PBS fidelity tools, such as the School-Wide Evaluation Tool (SET) (Horner, Todd, Lewis-Palmer, Irvin & Boland, 2004) and Benchmarks of Quality (BOQ) (Cohen, Kincaid, & Childs, 2007). These support evaluation and address the system's ongoing requirements.

Psychological basis of Positive Behaviour Support

PBS, including its use of functional assessment, 'is based directly on behavioural theory' and 'Applied Behavioural Analysis' (ABA) (Sugai & Horner, 2006, p. 247). ABA is grounded in radical behaviourism (Morris, Altus, & Smith, 2013). In application, behavioural modification is targeted by Skinnerian operant conditioning (Grindle et al., 2009), through behavioural reinforcement (Skinner, 1938). In ABA operant conditioning is adopted for behavioural modification, and the probability of appropriate behaviours are thought to increase through the use of immediate recognition and reinforcement. Problematic behaviours are expected to decrease when they are not reinforced (Grindle et al., 2009). The approach relies on principles similar to those described by Bronfenbrenner (1979), i.e. that behaviour is learned and governed by antecedent and consequential conditions. The emphasis on teaching and modelling behaviours in PBS also suggests a presence of social learning theory (Bandura, 1971).

Rationale for review

In England, data from the Department for Education (2019) suggests that school exclusions have been increasing since 2012. ‘Persistent disruptive behaviour’ is responsible for the majority of exclusions (Department for Education, 2019, p. 1). Significantly, exclusions result in missed classroom presence and therefore missed learning opportunities. Timpson (2019) stresses that every missed day of school affects a child’s chances of achieving. The terms ‘persistent’ and ‘disruptive’ suggest that challenging behaviours are detrimental to the learning of all pupils and contribute toward a negative educational experience.

PBS has been found to reduce noise levels in schools (Kartub et al., 2000), and to reduce disciplinary referrals (Bradshaw et al., 2015), suggesting that learning exposure can be maintained by reducing exclusions. PBS may demonstrate potential to increase learning focussed behaviour. Some studies have found it to be central to improving learning outcomes (Nelson, Martella & Marchand-Martella, 2002). However, its contribution to learning remains unclear. A systematic review conducted by Chitiyo et al., (2011) found a moderate effect on challenging behaviour and a questionable effect on learning attainment. A review by Gage et al. (2015) concluded that there is no relationship between PBS and academic achievement.

In the United States (US) PBS has been embraced in schools, both in specialist and mainstream settings. However, in the United Kingdom it is generally confined to few localities (Iemmi, Knapp, & Brown, 2016). The purpose of this review is to appraise PBS studies published since 2015 by

focusing on academic achievement. It aims to inform Educational Psychology practice, as the application and recommendation of school based interventions form a core function of an Educational Psychologist's role (Scottish Executive, 2002). The findings are synthesised in order to consider the suitability of PBS application in mainstream schools in the United Kingdom.

Review question

How effective is PBS at improving academic achievement in school aged children attending mainstream schools?

Critical review of the evidence

Literature Search

A systematic literature search was conducted on 16 December 2019. Searches were carried out using the following academic databases: *ERIC* (Education Resource Information Centre), *PsycINFO* and *Web of Science*. In total the search yielded 2014 results, after the removal of 22 duplicate results. These were screened by title and a further 1971 studies were removed. 43 articles were then screened by abstract, removing 23 studies. A further 7 studies were identified and located through ancestry and citation searching. After reviewing the full text of 27 papers, 20 papers were screened and 7 studies that met the inclusion criteria were selected for analysis. Table 2 outlines the inclusion and exclusion criteria used in the

screening process. Excluded studies are listed in Appendix A, Table 1.

The final studies are listed in Table 3.

Table 1

Search Terms used in ERIC, PsycINFO & Web of Science Searches

Intervention	Participants	Context	Outcome
“positive behav* support"	Child	School	Effect
“positive behaviour intervention”	Pupil		Academic achievement
“PBS”	student		
“SWPBS”			
SWPBIS”			
“PALS”			

the asterisk (*) is a truncation, used to search for different word endings. For example: behaviour, behavior, behave, behavioural etc.

Figure 1

Flowchart of the Literature Search and Screening Process

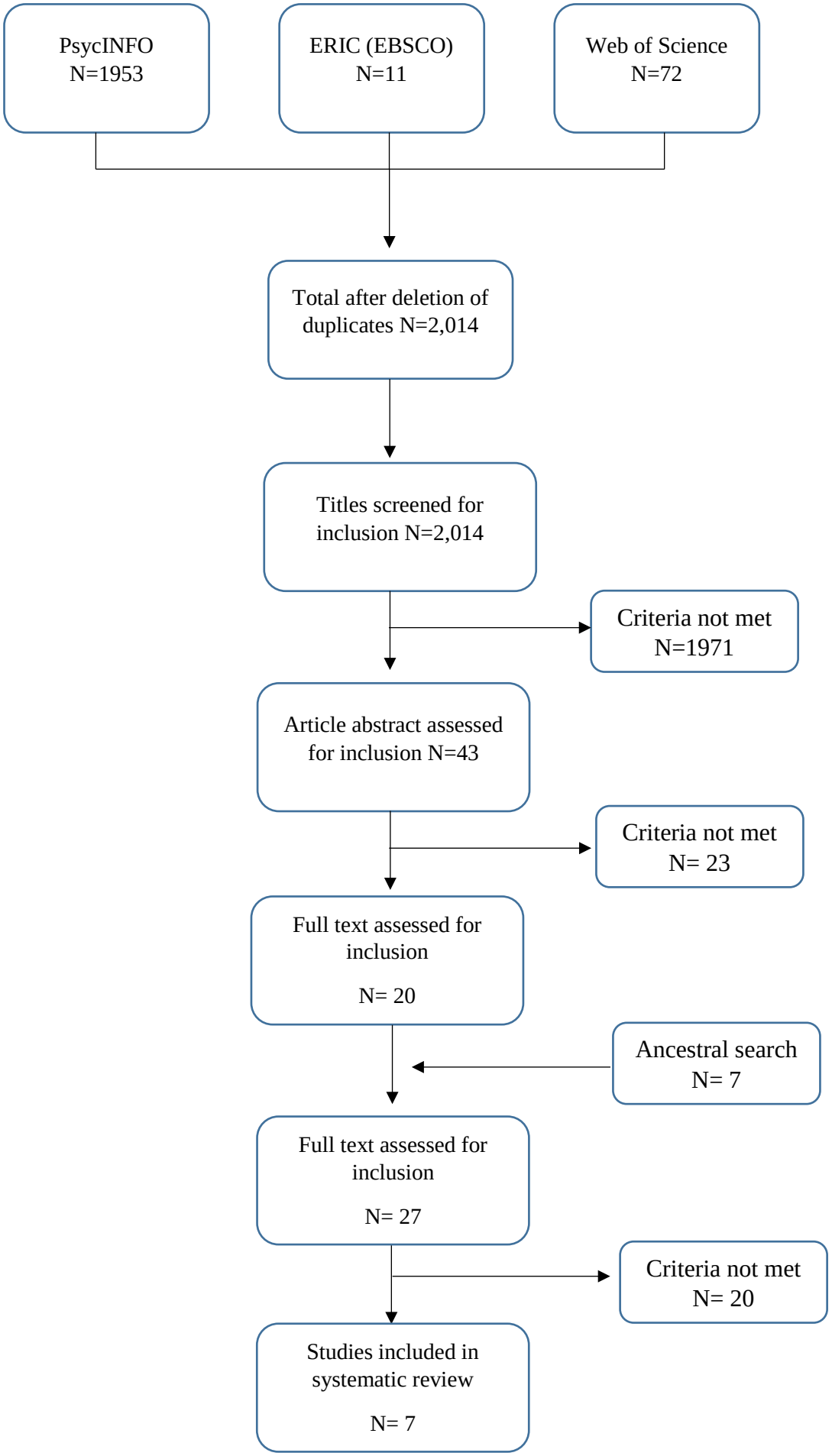

Table 2

Inclusion and Exclusion Criteria, (adapted from Spier et al., 2016)

	Inclusion criteria	**Exclusion criteria**	**Rationale**
1. Type and date of publication	Peer reviewed journal articles dated between 2015 and 2019	Non-peer reviewed journals, books, dissertation and grey literature	To ensure expert scrutiny of scholarly material A systematic review was completed by (Gage et al., 2015). This review aims to include the most recent studies that were not included in previous reviews
2. Population	School aged children from Reception Year to Year 11 (or country equivalent), ages 4 to 16 years	Children below the age of 4 years, young people above the age of 16 years. Individuals who are not attending a school setting	The aim of the review is to assess the effectiveness of the intervention when used in a school setting/context
3. Intervention	Positive Behaviour Intervention and Support programmes, delivered on a whole school basis, often referred to as School Wide PBS/PBIS	Any other behavioural models or interventions	The aim of this review is to explore the effect of the specific intervention when delivered to a whole school or class

	Inclusion criteria	Exclusion criteria	Rationale
4. Type of setting	School based intervention	Home based school programmes or studies conducted in clinical settings	The aim of the review is to explore the effectiveness of the intervention in a school context
5. Outcome measure	Quantitative measure of academic achievement, including baseline measurements A measure of intervention fidelity	Studies that only measure behavioural outcomes, such as exclusion or discipline	To measure the effect of the intervention on learning outcomes / academic achievement.
6. Geographic context	Studies from The Organisation for Economic Co-operation and Development (OECD) countries	Studies carried out in countries that are not part of the OECD	There are similarities between countries' policies and principles which impact on education and culture, making such studies more suitable for generalisability to the UK context
7. Research design	Empirical studies including: Randomized Control Trials (RCTs), Quasi –experimental studies, Cohort studies and longitudinal studies	Case studies, cross sectional studies. Descriptive studies and qualitative reports	Based on Petticrew and Roberts (2008) typology of suitable studies when reviewing effectiveness

	Inclusion criteria	Exclusion criteria	Rationale
		Systematic reviews and meta-analysis	
8. Language	Articles / studies published in English	Articles / studies not published in English	Reviewer able to access and comprehend the study. To ensure accurate content due to reliability concerns with online translation tools

Table 3

Final Studies Included in the Systematic Review

Angus, G., & Nelson, R. B. (2019). School-Wide Positive Behavior Interventions and Supports and Student Academic Achievement. *Contemporary School Psychology*, No-Specified.
Borgen, N. T., Kirkeboen, L. J., Ogden, T., Raaum, O., & Sorlie, M.-A. (2019). Impacts of school-wide positive behaviour support: Results from national longitudinal register data. *International Journal of Psychology*.
Freeman, J., Simonsen, B., McCoach, D. B., Sugai, G., Lombardi, A., & Horner, R. (2016). Relationship Between School-Wide Positive Behavior Interventions and Supports and Academic, Attendance, and Behavior Outcomes in High Schools. *Journal of Positive Behavior Interventions, 18*(1), 41–51.
Houchens, G. W., Zhang, J., Davis, K., Niu, C., Chon, K. H., & Miller, S. (2017). The impact of Positive Behavior Interventions and Supports on teachers' perceptions of teaching conditions and student achievement. *Journal of Positive Behavior Interventions*, *19*(3), 168–179.
Madigan, K., Cross, R. W., Smolkowski, K., & Strycker, L. A. (2016). Association between schoolwide positive behavioural interventions and supports and academic achievement: a 9-year evaluation. *Educational Research and Evaluation*, *22*(7–8), 402–421.
Reno, G. D., Friend, J., Caruthers, L., & Smith, D. (2017). Who's getting targeted for behavioral interventions? Exploring the connections between school culture, positive behavior support, and elementary student achievement. *Journal of Negro Education*, *86*(4), 423–438.
Ryoo, J. H., Hong, S., Bart, W. M., Shin, J., & Bradshaw, C. P. (2018). Investigating the effect of school-wide positive behavioral interventions and supports on student learning and behavioral problems in elementary and middle schools. *Psychology in the Schools*, *55*(6), 629–643.

Table 4

Mapping the Field

Study	Location	Participants	Intervention	Design	Measures	Outcomes
Angus & Nelson (2019)	Urban Southern California school district (United States of America - USA)	All year groups (grades 6, 7 & 8) in 8 schools with a selected pupil sample of 8515	School Wide Positive Behaviour Interventions and Support (SWPBIS)	Quasi-experimental, repeated treatment design, with longitudinal data	2003 to 2010 mean scores on California Standards Test (CST) for 8 Schools were compared to school archival data to see if PBS was related to effects on student attainment in English and maths. The relationship between implementation fidelity and achievement were also measured.	**English:** Overall, implementation with fidelity was strongly related to an increase in test scores for all three year groups. Changes in test scores from 2004 to 2010 were statistically significant: $F(1, 6) = 17.56$, $p = 0.0001$ **Maths:** PBS implementation was associated to increases in maths test scores. The increase in test scores over time were statistically significant (Greenhouse- Geisser analysis was used for correction $F(1, 2.77) = 5.90$, $p = 0.01$.

Study	Location	Participants	Intervention	Design	Measures	Outcomes
Borgen et al. (2019)	Norway	2365 Primary schools (grades 1–7)	N-PALS, School Wide Positive Behaviour Support (SWPBS)	Difference-in-difference (DiD) design, using longitudinal data	Primary measures included: classroom noise and bullying through self-report. Secondary measures included: pupil wellbeing, measured through self-report and academic attainment. These were measured through standardised national tests in literacy, English and numeracy. Register data comparison of PBS schools and control schools were analysed for 3 years before and 5 years after the intervention.	No intervention effect was observed on academic performance (based on scores on standardised national tests in literacy and numeracy ($p < .10$).

Study	Location	Participants	Intervention	Design	Measures	Outcomes
Freeman et al. (2016)	37 states across the USA	883 high schools in the intervention group. Data from 934 'no intervention' control schools were used to control for student exposure.	School-Wide Positive Behaviour Interventions and Supports (SWPBIS)	A quasi-experimental interrupted time series design, with longitudinal data	Implementation fidelity was measured using two tools. These were matched to suitable controls and confirmatory factor analysis (CFA) was used to explore fidelity consistency with school attendance data, disciplinary referrals, maths, reading and language index (school and state data). 2005-2012 data was measured.	Intervention fidelity was not related to academic outcomes in a statistically significant manner. Effects were negative (−0.13, $p = .69$) for schools implementing without fidelity and positive (0.26, $p = .54$) for schools that were implementing with fidelity. Neither were statistically different.

Study	Location	Participants	Intervention	Design	Measures	Outcomes
Houchens et al. (2017)	Kentucky, USA	151 school (95 elementary, 31 middle, 25 high) Control schools: 144 (89 elementary, 30 middle, 25 high)	School-Wide Positive Behavioural Interventions and Supports (SWPBIS)	A quasi-experimental design.	Implementation fidelity was measured using two assessments. Student academic scores and demographic information were measured, based on state records. Data from academic year 2010-2011 was analysed. The study also sought teacher perceptions of conditions in PBIS schools.	Overall - no significant differences in pupil attainment levels between the intervention and non-intervention (control) schools. A significant difference was found between different levels of intervention implementation on overall test scores, $F(3, 291) = 3.42$, $p = .01$. Post hoc multiple comparisons using the Bonferroni correction found that high and medium-fidelity PBS schools and intervention schools achieved significantly higher overall test scores than low-fidelity PBS schools ($p < .05$). No significant difference in test scores between high and medium fidelity PBS schools. No significant overall test score increase was found between treatment and no treatment groups – tested by aggregating low, medium, and high-fidelity PBS groups.

Study	Location	Participants	Intervention	Design	Measures	Outcomes
Madigan et al. (2016)	Kentucky, USA	21 schools (15 elementary, 5 middle, and 1 high). A total of 11,202 pupils. 28 comparable schools (21 elementary, 5 middle, and 2 high). A total of 14,857 pupils	Foundations, Positive Behaviour Interventions and Support	A quasi-experimental design with longitudinal data	The study examined the association between whole school PBS implementation and academic attainment. Comparison to historical index data and schools that use traditional behaviour management and discipline practices. State academic index data was analysed. Intervention fidelity was also measured.	Results indicate that PBS implementation was significantly associated with increased pupil attainment ($p = .001$). The rate of change for in achievement in treatment schools was greater than for students in 'no treatment control schools. $F8,38 = 4.35$, $p < .001$).

Study	Location	Participants	Intervention	Design	Measures	Outcomes
Ryoo et al. (2018)	Minnesota, USA	33 intervention schools and 33 (no intervention) control schools	School Wide Positive Behaviour Support (SW-PBIS)	A quasi-experimental design with longitudinal data	Minnesota Comprehensive Assessment (MCA) test scores from 2007–2008 to the 2009–2010 academic year. Implementation fidelity was also measured.	The study did not find any statistically significant effects of PBS on academic achievement.

Table 4b*

M*apping the Field*

Study	Location	Participants	Intervention	Design	Measures	Outcomes
Reno et al. (2017)	Unknown Midwest state, USA	71 pupils in the intervention group with 71 pupils in a (no intervention) control group	Positive Behaviour Intervention and Support (PBIS) Tier II intervention of Check-In/Check-Out (CICO)	A quasi-experimental design using randomized control	The study measured reading and maths levels using The Renaissance Learning STAR reading and math assessments. It assessed these levels against behavioural data collected for the intervention. Teacher self-report of pupil involvement and academic achievement were also used. Data represented the academic years 2011-2012 and 2012-2013.	No statistically significant change found between the intervention group and academic achievement.

* this study is reported separately, because it measured the impact of the Tier II version of the intervention on academic achievement

Weight of evidence

Included studies were evaluated using the Gough (2007) Weight of Evidence framework 'for the appraisal of the quality and relevance of evidence' (p. 221). The framework enables consistent and explicit evaluations regarding 'the quality of execution and appropriateness of design to answer the review question' (Gough, 2007, p. 222), by assigning a rating to three key dimensions. Weight of Evidence A (WoE A) provides a judgement of the methodology quality employed (Gough, 2007). Weight of Evidence B (WoE B) provides a judgement of the relevance of the methodology to the review question, and Weight of Evidence C (WoE C) - a judgement of the topic relevance in relation to the review question (Gough, 2007). WoE A, B and C are combined and an average is calculated for each study, producing an overall judgement score, known as Weight of Evidence D (WoE D) (Gough, 2007).

To assess WoE A, an adapted version of the *Task Force on Evidence-Based Interventions in School Psychology* (Kratochwill, 2003) coding protocol for group designs was employed. WoE A descriptors, criteria, mean score range, overall scores and amendments are located in Appendix B. A copy of a WoE A protocol is located in Appendix C. WoE B is based on Petticrew and Roberts' (2003) 'example of typology of evidence' criteria, located in Appendix D. A criteria for WoE C is located in Appendix E. A summary of all weight of evidence scores is located in Appendix F.

Table 5
Summary of Weight of Evidence

Study	WoE A	WoE B	WoE C	WoE D
Angus & Nelson (2019)	1.5 (Medium)	1 (Low)	3 (High)	1.83 (Medium)
Borgen et al. (2019)	0.66 (Low)	2 (Medium)	2 (Medium)	1.55 (Medium)
Freeman et al. (2016)	1.3 (Low)	2 (Medium)	1 (Low)	1.43 (Low)
Houchens et al., (2017)	2.5 (High)	3 (High)	2 (Medium)	2.5 (High)
Madigan et al. (2016)	2.16 (Medium)	3 (High)	3 (High)	2.72 (High)
Reno et al. (2017)	1.66 (Medium)	2 (Medium)	1 (Low)	1.55 (Medium)
Ryoo et al. (2018)	2.0 (Medium)	2 (Medium)	3 (High)	2.3 (Medium)

*WoE D rating descriptors: 'High' ≥2.5, 'Medium' 1.5 –2.4, 'Low' ≤1.4

Participants

The studies in this review were conducted in Norway and in the US. They included primary, middle and high schools and data from more than 39,500 pupils. Sample sizes ranged from one school with a group of 71 elementary age pupils, to 883 high schools across 37 states.

Reno et al. (2017) considered the pupils to be participants, whereas all other studies considered the schools themselves to be the participating bodies, due to experimental size. As a result of the large samples, significant pupil information, such as gender, were omitted. This limitation was recognised by Houchens et al. (2017) who provided information regarding educators and sought additional self-report data.

Each study reported pupil year groups/grades and demographic information relating to social economic status. These variables were used for the assignment of suitable control groups in four studies (Freeman et al., 2016;

Houchens et al., 2017; Madigan et al., 2016; Ryoo et al., 2017), by matching characteristics that can influence achievement. Sample sizes were taken into account in WoE A. Power calculations suggest that all of the studies were sufficiently powered. The sampling method and diversity of samples i.e. representative for wider generalisation, were considered in the criteria for WoE A and B, respectively.

Most of the studies based sampling and inclusion on intervention fidelity measures. WoE B reviewed the robustness of the fidelity measures used in each study. Freeman et al. (2016) and Ryoo et al. (2017) selected schools that were implementing PBS and had reached fidelity according to the SET (Horner, et al., 2004) and the BOQ tool (Cohen, Kincaid & Childs, 2007). Houchens et al. (2017) grouped schools based on fidelity scores according to the BOQ. Madigan et al. (2016) included schools that had achieved ‘medium’ to ‘high’ fidelity. Unlike other studies which used fidelity measures to determine participant inclusion, convenience sampling was adopted by Angus and Nelson (2019). These schools were implementing PBS under district mandate, indicating potential issues of bias or poor generalizability (Barker, Pinstrang & Elliott, 2016). Borgen et al. (2019) included all primary schools in Norway from grades 1-7. More uniquely, Reno et al. (2017) used criterion sampling to identify pupils targeted for Tier II PBS, after selecting a school for durational use of PBS.

Study Design

All of the studies in this review were quasi-experimental designs using longitudinal data. A longitudinal approach was considered to be suitable because PBS methods need to be integrated into school culture (Sugai et al., 2000). Control groups receiving 'no intervention' were used by Houchens et al. (2017); Madigan et al. (2016); Reno et al. (2017) and Ryoo et al. (2018). Angus and Nelson (2019); Borgen et al. (2019) and Freeman et al. (2016) used variables from school data as covariates to control for changes following intervention implementation. Those with control groups received a higher WoE A than those that relied on statistical controls. Reasoning is based on Cook, Campbell and Shadish's (2002) suggestion that research has not:

> '…much supported the use of statistical adjustments in longitudinal national surveys in which individuals with different experiences are explicitly contrasted in order to estimate the effects of this experience difference' (Cook, Campbell & Shadish, 2002 p.504).

In such situations 'undermatching is a chronic problem' due to 'consequences of unreliability in the selection variables' and specification errors are possible 'due to incomplete knowledge of the selection process' (Cook, Campbell & Shadish, 2002 p.504).

Houchens et al. (2017) applied propensity score matching (Guo & Fraser, 2015) to reduce selection bias of unbalanced data. A high WoE B was assigned to Houchens et al. (2017) and Madigan et al. (2016) as both constructed control groups by matching participants (Rossi et al., 2004),

instead of adopting a one-group pretest-posttest design which has associated threats to internal validity (Cook & Campbell, 1979). Ryoo et al. (2018) also applied the propensity score matching method, but only used two academic subjects to measure impact.

Measures

WoE A considered the reliability of measures, including stable baseline measurements. Most commendable for use of reliable baseline measurements were: Angus and Nelson (2019); Borgen et al. (2019); Freeman et al. (2016); and Madigan et al. (2016), who measured school exam results for a number of years before PBS introduction. Madigan et al. (2016) confirmed that there were no confounding interventions in treatment schools, which might otherwise have skewed results. Houchens et al. (2017) sought additional insight in the form of teachers' perceptions on the impact of PBS. Despite stable baseline measurements, WoE A scores for Borgen et al. (2019) were low because academic progress was a secondary focus of the research, therefore lacking stringent review. Not all data was available for all year groups, schools or subjects in the Freeman et al. (2016) study.

Studies that measured the most academic subjects benefited from increased reliability of measurement, as individual pupil differences were reduced. These offered a wider perspective of change in learning performance. The type of test data used to analyse progress was fundamental to WoE C. Criterion based examinations, such as State tests and national tests offer a high level of reliability. They have better face validity than standardised or

norm referenced tests because they are grounded in the school curriculum. For instance, the California State Test used by Angus and Nelson (2019) has a high internal consistency of α=0.94. Madigan et al. (2016) reviewed multiple subjects using data from state examinations. The highest scoring studies measured academic attainment across multiple cohorts using criterion referenced state tests, across a longitudinal time series. The lowest scores were allocated to Freeman et al. (2016) who lacked a test of heterogeneity across the sample population, and Reno et al. (2017) who based progress on a norm referenced assessment. These offered lower validity and raise concern regarding participant practice effects.

Fidelity

Consideration of implementation fidelity was central to the WoE B. The most stringent approaches reduced the likeliness of a regression toward the mean. Fidelity is critical because learning can only improve if PBS first reduces disruptive behaviour (Ryoo et al., 2018).

Angus and Nelson (2019) reported that schools were only considered to be implementing when all of the components in the PBS framework had been gathered by an external coach. Schools had dedicated PBS teams and quarterly meetings with a coach. To ensure consistency of application and self-report scores, the coach met with senior staff and analysed data. Houchens et al. (2017) categorised high, medium and low fidelity for analysis, using the widely recognised BOQ, which has a high internal consistency of .96, calculated using Cronbach coefficient alpha (Cohen,

Kincaid & Childs, 2007, cited by Houchens et al., 2017). Madigan et al. (2016) only included schools that had achieved high fidelity according to a Self-Assessment Survey (Sugai, Horner, & Todd, 2003). Most rigorous, was the study by Ryoo et al. (2018), which required schools to use the BOQ through self-report. It was stressed that schools could only use self-report once a high score had been achieved according to the SET. All schools completed two years of training, minimising intervention variance.

Fidelity approaches were weaker in the Freeman et al. (2016) study. Schools with 'low fidelity' were classified equally to those 'not implementing', compromising the ability to detect smaller changes in academic performance between groups. Analysis of the long term effect of PBS outcomes were not possible, as the 'number of secondary schools that had implemented PBS with fidelity for more than two years was limited' (Freeman et al., 2016, p. 43). The study's large national scale and reliance on self-report meant that 9% of fidelity data was missing. Borgen et al. (2019) used annual self-report data which strongly threatened fidelity. Reno et al. (2018) recognised concerns regarding inter-rate reliability of fidelity ratings.

Intervention

It is predictable that approaches to PBS will differ by varying degrees in all of the schools. This is because each school identifies specific behaviours that are problematic, for modification. Madigan et al. (2016) explain that PBS

> '…is not a prescriptive programme, but a structured process to help schools make decisions about discipline needs. Therefore, schools' disciplinary and school climate outcomes may vary, depending on their

identified and prioritized needs, and the time and resources devoted…' (p. 411).

Nonetheless, intervention similarities are strongly implied in each study, with highly consistent themes noted in all treatment schools. Fidelity measurement tools (PBIS.org Assessments, 2020) all contain consistent themes. The study of PBS application in Norway (Borgen et al., 2019) describes PBS components and implementation as equal to the U.S. counterpart.

A critical feature of all schools is the quality of teaching and learning opportunities. None of the studies report on teaching quality or the academic support offered in treatment schools, which may be significant to the intervention impact on the dependant variable.

Outcome and effect sizes

Mixed outcomes of the effectiveness of PBS at improving academic achievement in school aged children were found. The majority of studies did not observe a statistically significant effect between PBS and academic achievement. Each study used a quantitative approach to measure correlation. Reported calculations (Angus & Nelson, 2019; Houchens et al., 2017; Madigan et al., 2016) were converted to Cohen's d for comparison, using the Campbell Collaboration effect size calculator (Wilson, 2020). Effect sizes (Table 6) were calculated using reported means and standard deviations for the studies that did not report effect sizes.

A large effect on both English and maths was reported by Angus and Nelson (2019), which was statistically significant. The study did not specify the name of the effect size used. A calculation of Cohen's d was therefore completed, which suggested a medium effect on English and maths. Similarly, Madigan et al. (2016) found a large effect, using η_p^2 and Hedges g. The authors concluded that PBS was significantly associated with increased student academic achievement. The study established that the rate of academic achievement change in treatment schools was greater than for students attending control schools.

Borgen et al. (2019); Freeman et al. (2016); Ryoo et al. (2018) found no statistically significant intervention effect on performance. Comparably, Reno et al. (2017) did not find a statistically significant connection between intervention participation and academic achievement. This finding should be considered in isolation, since it only suggests that targeted PBS does not improve academic achievement in comparison to universal PBS. Notably, Houchens et al. (2017) did not find significant differences in attainment levels between intervention and control schools. However, analysis revealed that outcomes were significantly better in high and medium fidelity schools than schools implementing PBS with low fidelity (Houchens et al., 2017).

Table 6

Effect Sizes and Descriptors

Study	Sample size and groups	Outcome and Significance	Effect size (Cohen's d)	Description*	WoE D
Angus & Nelson (2019)	8 middle schools (8815 students). 0 control groups	PBS effect on English-Language Arts California Standards Test (CST): Statistically significant changes between 2004 and 2010 in test scores were found: F(1, 6) = 17.56, p = 0.0001'.	D = 0.71	Medium	Medium
		PBS effect on maths California Standards Test (CST): The increase in test scores were statistically significant (Greenhouse- Geisser analysis was used for correction F(1, 2.77) = 5.90, p = 0.01.	D = 0.70	Medium	
Borgen et al. (2019)	2365 primary schools. Comparing the 'post period' (to 'pre period' (DiD design)	Researchers did not observe an intervention effect on academic performance (scores on standardised national tests in literacy and numeracy were analysed ($p < .10$).	D = 0.012 (1-5 years of intervention)	Small	Low

Study	Sample size and groups	Outcome and Significance	Effect size (Cohen's d)	Description*	WoE D
			D = 0.013 (2-5 years)		
Freeman et al. (2016)	883 high (secondary) schools with 934 ('no intervention control') middle schools used in the analysis to control for student exposure to the intervention	Intervention fidelity scores were not related to learning outcomes in a statistically significant way. Effects were negative (−0.13, p = .69) for schools that had not reached fidelity. They were positive (0.26, p = .54) for schools that were implementing with fidelity. However, neither were statistically different from 0.	For high fidelity of intervention (Fidelity 1) D = 0.0132 For 'middle' fidelity of intervention (Fidelity 2) D = 0.0243	Small	Low
Houchens et al. (2017)	151 schools with 151 (no intervention) control schools	No significant differences found in student achievement levels between the intervention and no intervention (control) schools. However, a significant differences among different levels of intervention implementation was recorded on overall test scores, $F(3, 291) = 3.42$, p = .01	D = 0.0033 between intervention and non-intervention schools. (D = 0.1116	Small	High

Study	Sample size and groups	Outcome and Significance	Effect size (Cohen's d)	Description*	WoE D
		Post hoc multiple comparisons with Bonferroni correction identified that high-and medium-fidelity schools achieved significantly higher overall scores than low-fidelity PBS schools ($p < .05$). There was no significant difference in achievement scores was between high and medium fidelity PBS schools. Researchers aggregated low, medium, and high-fidelity schools, but found no significant overall score difference between PBS and no PBS schools. Surveys of teachers identified that high fidelity PBS was related to positive perceptions of teaching conditions.	between high fidelity schools and non-intervention schools) (D=0.189 between medium fidelity schools and non-intervention schools) (D=0 between low fidelity schools and non-intervention schools)		
Madigan et al. (2016)	21 schools with 28 comparable ('no intervention') control schools	Results showed that PBS implementation was significantly associated to increased academic achievement (p = .001). The rate of change for student achievement in PBS schools was greater than students in 'no treatment (control) settings. F8,38 = 4.35, p < .001).	D = 0.9051	Large	High

Study	Sample size and groups	Outcome and Significance	Effect size (Cohen's d)	Description*	WoE D
Ryoo et al. (2018)	33 intervention school and 33 ('no intervention') control schools	The study did not identify any statistically significant long term effect of PBS on academic achievement.	D = 0.2043	Small	Medium

The following study is reported separately, because unlike all other studies it measured the impact of the Tier II version of the intervention of academic achievement.

Study	Sample size and groups	Outcome and Significance	Effect size (Cohen's d)	Description*	WoE D
Reno et al. (2017)	A group of 71 pupils in the intervention group with 71 pupils in a (no intervention) control group	The study examined the impact of the Tier 2 (group) intervention. It did not find a statistically significant correlation between PBS participation and academic achievement (reading and maths).	D = 0.001 (Reading) D = 0.0009 (Maths)	Small	Medium

*descriptors from Watson (2019) citing Cohen (1988): ≥0.2 (small), ≥0.5 (medium), ≥0.8 (large)

The overall WoE D needs to be taken into account when considering the effect sizes of each study. Four individual studies ranging from low, medium and high WoE D scores found no statistically significant intervention effect on performance. These small effects ranged from d=0.0033 to d= 0.2043. In contrast, one study with a medium WoE D (Angus & Nelson, 2019) identified a medium effect size (d=0.71). The Madigan et al. (2016) study which received a high WoE D score identified a large effect (d=0.9051), demonstrating a scattered range of methodological quality and effect.

Each of the studies included in this review were appraised fairly, using Gough's Weight of Evidence framework (2007) with even weighting across WoE A, B and C. The framework enabled consistent evaluations to address the review question (Gough, 2007, p. 222). Table 6 shows that studies with a high methodological quality (Houchens et al., 2017; Madigan et al., 2016) found a small insignificant effect and a large effect, respectively, in primary outcomes. This demonstrates that studies with significant findings were not more highly weighted than those without significant findings in this review. Weight of Evidence ratings are taken into account in the concluding interpretation, which suggests mixed evidence for the effectiveness of PBS at improving academic achievement. Furthermore, Table 5 shows that two studies received a higher WoE C rating than the Houchens et al (2017) study. Table 5 also presents a number of studies that received the same overall WoE A and C rating as the two highest scoring (WoE D) studies.

Conclusions and recommendations

The objective of this systematic literature review was to examine the effect of PBS on academic achievement in school aged children attending mainstream schools. It aimed to consider intervention benefits for generalisation and application in the context of the United Kingdom where exclusions have been increasing since 2012 (Department for Education, 2019). Using Gough's (2007) weight of evidence framework, this review contributes toward existing literature, (Chitiyo et al., 2011; Gage et al., 2015) through the evaluation of PBS studies published between 2016 and 2019.

This review found mixed evidence for the effectiveness of PBS at improving academic achievement. In line with the conclusion of Gage et al. (2015), four studies found no intervention effect on academic performance for pupils ranging through elementary, primary and high school. Additionally, Reno et al. (2017) concluded that there was no statistically significant connection between the Tier 1 and Tier II interventions, for primary aged pupils. Among these publications was one study with a 'high' overall WoE D (Houchens et al., 2017). Three of the studies obtained a 'medium' WoE D (Borgen et al., 2019; Reno et al., 2017; Ryoo et al., 2018) and one received a 'low' WoE D (Freeman et al., 2016).

Two of the studies identified strong relationships between PBS and student achievement. The first found a medium effect size on increases in test scores for maths and English, with improvements increasing in parallel to

intervention fidelity (Angus & Nelson, 2019). This study received a ‘medium’ overall weight of evidence. Madigan et al. (2016) found that PBS was significantly associated with increased student academic achievement across multiple school subjects, with a large effect size. Notably, the study took place in the same US State as that conducted by Houchens et al. (2017). Although it is unknown if any schools were involved in both studies, this finding is anecdotal of the mixed results within the PBS literature.

Insightful conclusions can be drawn from Ryoo et al. (2018) who recognise that schools may not have implemented PBS as it had been intended. This observation is transferable to a number of low scoring fidelity approaches in this review, which found no impact on learning achievement. Houchens et al. (2017) identified that schools implementing PBS with high and medium fidelity achieved higher examination scores than those implementing with low fidelity. ‘High’ fidelity implementation was associated with positive perceptions of teaching conditions among teachers. These findings strongly imply that schools considering the implementation of PBS should do so with the highest intention of achieving fidelity. Poor fidelity and/or inconsistent PBS may be confusing for students and possibly counterproductive.

Furthermore, PBS cannot improve academic performance if its application fails to reduce challenging and disruptive behaviour (Ryoo et al., 2018). Whilst PBS has been found to reduce challenging behaviours in a randomized control trial (Bradshaw et al., 2010), it is apparent that academic

achievement remains dependent on effective teaching approaches that are applied as a counterpart. Quality of teaching and a range of protective factors remain significant to learning. It is evident from high quality studies (Madigan et al., 2016; Houchens et al., 2017) that for PBS to improve learning, it needs to be applied with high fidelity, and should be complimented by high quality teaching that addresses the needs of individual learners.

Limitations and recommendations

The large samples of school level data in this review offered a broad view of the PBS impact through a range of ages spanning two countries. Large sample studies typically benefit from a high level of accuracy of mean data (Barker et al., 2016). However, sample size impacted the research potential to recognise specific groups of individual pupils that may benefit from PBS. Furthermore, lack of random assignment of treatment schools was a weakness in all of the included studies, restricting their ability to assert a causal relationship (Madigan et al., 2016). None of the studies reported levels of special educational needs (SEN), which is significant because school levels of inclusion may potentially alter mean academic attainment scores.

It is recommended that future research focuses on interventions that can be used in conjunction with PBS to explore the potential of a ‘pincer approach’ on learning and achievement. Research should focus on the impact that PBS may have for specific subgroups of pupils in schools. Such approaches

should use randomized control trials to seek epidemiologic evidence (Barker et al, 2016) incorporating key principles of the Bradford Hill criteria (Schünemann et al., 2011) to determine causality.

The literature reviewed in this systematic study suggests a lack of pupil voice in relation to experiences with PBS. Mixed method designs should be considered in order to seek qualitative views from pupils involved with PBS and quantitative measures of subsequent progress. Research in this area, could offer a richer description of the changes that PBS offers to the learning experience.

References

American Psychological Association. (2010). *Publication manual of the American Psychological Association (6th ed*.). Washington, D.C.: American Psychological Association.

Angus, G., & Nelson, R. B. (2019). School-Wide Positive Behavior Interventions and Supports and Student Academic Achievement. *Contemporary School Psychology*. https://doi.org/10.1007/s40688-019-00245-0

Barker, C., Pinstrang, N. & Elliott, R. (2016). *Research Methods in Clinical Psychology, An Introduction for Students and Practitioners* (3rd ed.). Oxford: Wiley Blackwell.

Bandura, A. (1971). *Social Learning Theory*. New York, NY: General Learning Press.

Beaman, R., & Wheldall, K. (2000). Teachers' Use of Approval and Disapproval in the Classroom. *Educational Psychology*, 20(4), 431–446. https://doi.org/10.1080/713663753

Borgen, N. T., Kirkeboen, L. J., Ogden, T., Raaum, O., & Sorlie, M.-A. (2019). Impacts of school-wide positive behaviour support: Results from national longitudinal register data. *International Journal of Psychology*, No-Specified. https://doi.org/http://dx.doi.org/10.1002/ijop.12575

Bradshaw, C. P., Mitchell, M. M., & Leaf, P. J. (2010). Examining the effects of schoolwide positive behavioral interventions and supports on student outcomes: Results from a randomized controlled effectiveness trial in elementary schools. *Journal of Positive Behavior Interventions*, *12*(3), 133–148.

Bradshaw, C. P., Waasdorp, T. E., & Leaf, P. J. (2015). Examining variation in the impact of school-wide positive behavioral interventions and supports: Findings from a randomized controlled effectiveness trial. *Journal of Educational Psychology*, 107(2), 546–557. https://doi.org/10.1037/a0037630

Bronfenbrenner, U. (1979). *The ecology of human development*. Harvard university press.

Chitiyo, M., Makweche-Chitiyo, P., Park, M., Ametepee, L. K., & Chitiyo, J. (2011). Examining the effect of positive behaviour support on academic achievement of students with disabilities. *Journal of Research in Special Educational Needs*, 11(3), 171–177. https://doi.org/http://dx.doi.org/10.1111/j.1471-3802.2010.01156.x

Chu, S.-Y. (2015). An Investigation of the Effectiveness of Family-Centred Positive Behaviour Support of Young Children with Disabilities. *International Journal of Early Years Education*, 23(2), 172–191. Retrieved from:

http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=eric&AN=EJ1069168&site=ehost-live&scope=site

Cohen, J (1988) *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, R., Kincaid, D., & Childs, K. E. (2007). Measuring school-wide positive behavior support implementation: Development and validation of the benchmarks of quality. *Journal of Positive Behavior Interventions*, 9(4), 203–213. https://doi.org/10.1177/10983007070090040301

Cook, T.D., & Campbell, D.T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Chicago: Rand-McNally.

Cook, T. D., Campbell, D. T., & Shadish, W. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Department for Education. (2019). *Permanent and fixed period exclusions in England: 2017 to 2018*. (July). Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/820773/Permanent_and_fixed_period_exclusions_2017_to_2018_-_main_text.pdf

Freeman, J., Simonsen, B., McCoach, D. B., Sugai, G., Lombardi, A., & Horner, R. (2016). Relationship Between School-Wide Positive Behavior Interventions and Supports and Academic, Attendance, and Behavior Outcomes in High Schools. *Journal of Positive Behavior Interventions*, 18(1), 41–51. https://doi.org/10.1177/1098300715580992

Gage, N. A., Sugai, G., Lewis, T. J., & Brzozowy, S. (2015). Academic Achievement and School-Wide Positive Behavior Supports. *JOURNAL OF DISABILITY POLICY STUDIES, 25(4), 199–209. https://doi.org/10.1177/1044207313505647*

Gagnon, J. C., Rockwell, S., Scott, T. M. (2008). Positive behavior supports in exclusionary schools: A practical approach based on what we know. *Focus on Exceptional Children*, 41(1), 1–20.

Gough, D. (2007). *Weight of evidence:* A framework for the appraisal of the quality and relevance of evidence*. Research Papers in Education*, 22(2), 213–228. https://doi.org/10.1080/02671520701296189

Grindle, C., Hastings, R., Saville, M., Hughes, J.C., Kovshoff, H., & Huxley, K. (2009). Integrating evidence-based behavioral teaching methods into education for children with autism. *Educational and Child Psychology*. 26(4). 65-81.
Guo, S., & Fraser, M.W. (2015). *Propensity score analysis: Statistical methods and applications*. (2nd ed.) ThousandOaks, CA: Sage.

Horner, R. H., Todd, A. W., Lewis-palmer, T., Irvin, L. K., & Boland, J. B. (2004). The School-Wide Evaluation Tool (SET)*: 6*(1), 3–12.

Houchens, G. W., Zhang, J., Davis, K., Niu, C., Chon, K. H., & Miller, S. (2017). The impact of Positive Behavior Interventions and Supports on teachers' perceptions of teaching conditions and student achievement. *Journal of Positive Behavior Interventions*, 19(3), 168–179. https://doi.org/http://dx.doi.org/10.1177/1098300717696938

Iemmi, V., Knapp, M., & Brown, F. J. (2016). Positive behavioural support in schools for children and adolescents with intellectual disabilities whose behaviour challenges: An exploration of the economic case. *Journal of Intellectual Disabilities*, 20(3), 281–295. https://doi.org/http://dx.doi.org/10.1177/1744629516632402

Kartub, D. T., Taylor-Greene, S., March, R. E., & Horner, R. H. (2000). Reducing Hallway Noise: A Systems Approach. *Journal of Positive Behavior Interventions*, 2(3), 179–182. https://doi.org/10.1177/109830070000200307

Kelm, J. L., McIntosh, K., & Cooley, S. (2014). Effects of implementing school-wide positive behavioural interventions and supports on problem behaviour and academic achievement in a Canadian elementary school*. Canadian Journal of School Psychology*, 29(3), 195–212. https://doi.org/http://dx.doi.org/10.1177/0829573514540266

Kratochwill,T. R. (2003). Task Force on Evidence-Based Interventions in School Psychology. Retrieved 20 December, 2019, from http://www.indiana.edu/~ebi/documents/_workingfiles/EBImanual1.pdf

Lewis, T., & Sugai, G. (1999). Effective Behavior Support: A Systems Approach to Proactive Schoolwide Management*. Focus on Exceptional Children*. 31. 10.17161/foec.v31i6.6767.

Madigan, K., Cross, R. W., Smolkowski, K., & Strycker, L. A. (2016). Association between schoolwide positive behavioural interventions and supports and academic achievement: a 9-year evaluation*. Educational Research and Evaluation*, 22(7–8), 402–421. https://doi.org/10.1080/13803611.2016.1256783

Morris, E. K., Altus, D. E., & Smith, N. G. (2013). A study in the founding of applied behavior analysis through its publications. *Behavior Analyst*, 36(1), 73–107. https://doi.org/10.1007/BF03392293

Nelson, J.R., Martella, R.M., & Marchand-Martella, N. (2002). Maximizing Student Learning: The Effects of a Comprehensive School-Based Program for Preventing Problem Behaviors. *Journal of Emotional and Behavioral Disorders*, 10(3), 136–148. https://doi.org/10.1177/10634266020100030201

PBIS.org Assessments. (2020). Retrieved 9 February 2020, from https://www.pbis.org/resource-type/assessments

Petticrew, M., & Roberts, H. (2008). *Systematic Reviews in the Social Sciences: A Practical Guide*. In M. Petticrew & H. Roberts (Eds.), Systematic Reviews in the Social Sciences: A Practical Guide. https://doi.org/10.1002/9780470754887

Reno, G. D., Friend, J., Caruthers, L., & Smith, D. (2017). Who's getting targeted for behavioral interventions? Exploring the connections between school culture, positive behavior support, and elementary student achievement. *Journal of Negro Education*, *86*(4), 423–438. https://doi.org/10.7709/jnegroeducation.86.4.0423

Rossi, P.H., Lipsey, M.W., & Freeman, H.E. (2004) *Evaluation: A Systemic Approach* (7th ed.). Thousand Oaks, CA: Sage.

Ryoo, J. H., Hong, S., Bart, W. M., Shin, J., & Bradshaw, C. P. (2018). Investigating the effect of school-wide positive behavioral interventions and supports on student learning and behavioral problems in elementary and middle schools. *Psychology in the Schools*, 55(6), 629–643. https://doi.org/10.1002/pits.22134

Scottish Executive (2002). *Review of the Provision of Educational Psychology Services in Scotland.* Edinburgh: The Stationary Office.

Schünemann, H., Hill, S., Guyatt, G., Akl, E., & Ahmed, F. (2011). The GRADE approach and Bradford Hill's criteria for causation. *Journal of Epidemiology and Community Health (1979-), 65*(5), 392-395. Retrieved from http://www.jstor.org/stable/41150991.

Simonsen, B., Sugai, G., & Negron, M. (2008). Schoolwide Positive Behavior Supports: Primary Systems and Practices. *TEACHING Exceptional Children*, 40(6), 32–40. https://doi.org/10.1177/004005990804000604

Skinner, B.F. (1938).T*he Behaviour of Organisms*. New York: Appleton-Century.

Spier, E., Britto, P., Pigott, T., Roehlkapartain, E., McCarthy, M., Kidron, Y., Glover, J. (2016). Parental, Community, and Familial Support Interventions to Improve Children's Literacy in Developing Countries: A Systematic Review. *Campbell Systematic Reviews*, 12(1), 1–98. https://doi.org/10.4073/csr.2016.4

Sugai, G. & Horner, R. (2002) The Evolution of Discipline Practices: School-Wide Positive Behavior Supports. *Child & Family Behavior Therapy*, 24:1-2, 23-50, DOI: 10.1300/J019v24n01_03

Sugai, G., & Horner, R. H. (2009). Defining and describing schoolwide positive behavior support. In W. Sailor, G. Dunlop, G. Sugai, & R. Horner

(Eds.), *Issues in clinical child psychology. Handbook of positive behavior support* (p. 307–326). Springer Publishing Co. https://doi.org/10.1007/978-0-387-09632-2_13

Sugai, G., Horner, R. H., Dunlap, G., Hieneman, M., Lewis, T. J., Nelson, C. M., … Ruef, M. (2000). Applying Positive Behavior Support and Functional Behavioral Assessment in Schools. *Journal of Positive Behavior Interventions*, 2(3), 131–143. https://doi.org/10.1177/109830070000200302

Sugai, G., & Horner, R. R. (2006). A promising approach for expanding and sustaining School-wide positive behavior support. *School Psychology Review*, 35(2), 245–259.

Sugai, G., Horner, R., Sailor, W., Dunlap, G., Eber, L., & Lewis, T. (2004). School-wide positive behavior support: Implementers' blueprint and self-assessment. *University of Oregon, Center on Positive Behavioral Interventions and Supports, Eugene. Retrieved from Http://Pbis. Org/English/Handouts. Htm On*, *8*, 2004.

Sugai, G., Horner, R., & Todd, A. (2003). EBS Self-Assessment Survey, Version 2.0. Eugene, OR: Educational and Community Supports, University of Oregon.

Timpson, E. (2019). *Timpson Review of School Exclusion*. Retrieved from https://www.gov.uk/government/consultations/school-exclusions-review-call-for-evidence

Watson, P. (2019, October 30) Rules of thumb on magnitudes of effect sizes. Retrieved from: https://www.mendeley.com/guides/apa-citation-guide

Wilson, D. B., (2020). Practical Meta-Analysis Effect Size Calculator [Online calculator]. Retrieved: 09 February 2020, from: https://campbellcollaboration.org/research-resources/effect-size-calculator.html

Appendices

Appendix A

Table 1

List of Excluded Studies

Study	Reason for exclusion
Bradshaw, C. P., Debnam, K., Koth, C. W., & Leaf, P. (2009). Preliminary validation of the implementation phases inventory for assessing fidelity of schoolwide positive behavior supports. *Journal of Positive Behavior Interventions*, *11*(3), 145–160.	1, 5
Bradshaw, C. P., Mitchell, M. M., & Leaf, P. J. (2010). Examining the effects of schoolwide positive behavioral interventions and supports on student outcomes: Results from a randomized controlled effectiveness trial in elementary schools. *Journal of Positive Behavior Interventions*, *12*(3), 133–148.	1, 5,
Chu, S.-Y. (2015). An Investigation of the Effectiveness of Family-Centred Positive Behaviour Support of Young Children with Disabilities. *International Journal of Early Years Education*, *23*(2), 172–191.	5, 6, 7
Gage, N. A., Sugai, G., Lewis, T. J., & Brzozowy, S. (2015). Academic Achievement and School-Wide Positive Behavior Supports. *JOURNAL OF DISABILITY POLICY STUDIES*, *25*(4), 199–209.	7
Hill, D., & Brown, D. (2013). Supporting Inclusion of At Risk Students in Secondary School through Positive Behaviour Support. *International Journal of Inclusive Education*, *17*(8), 868–881.	1, 7
Horner, R. H., Sugai, G., Smolkowski, K., Eber, L., Nakasato, J., & Todd, A. W. (2009). Support in Elementary Schools. *Journal of Positive Behavior Interventions*, *11*, 133–144.	1
Iemmi, V., Knapp, M., & Brown, F. J. (2016). Positive behavioural support in schools for children and adolescents with intellectual disabilities whose behaviour challenges: An exploration of the economic case. *Journal of Intellectual Disabilities*, *20*(3), 281–295.	5
Lane, K.L., and Menzies, H.M,. (2019). A School-Wide Intervention with Primary and Secondary Levels of Support for Elementary Students : Outcomes and Considerations. *Education and Treatment of Children , Vol . 26 , No . 4 (NOV*. *26*(4), 431–451.	7

Study	Reason for exclusion
Kelm, J. L., McIntosh, K., & Cooley, S. (2014). Effects of implementing school-wide positive behavioural interventions and supports on problem behaviour and academic achievement in a Canadian elementary school. *Canadian Journal of School Psychology*, *29*(3), 195–212.	1, 7
Lassen, S. R., Steele, M. M., & Sailor, W. (2006). The Relationship of School-Wide Positive Behavior Support to Academic Achievement in an Urban Middle School. *Psychology in the Schools*, *43*(6), 701–712.	1
Luiselli, J. K., Putnam, R. F., Handler, M. W., & Feinberg, A. B. (2005). Whole-School Positive Behaviour Support: Effects on student discipline problems and academic performance. *Educational Psychology*, *25*(2–3), 183–198.	1
McIntosh, K., Bennett, J. L., & Price, K. (2011). Evaluation of Social and Academic Effects of School-Wide Positive Behaviour Support in a Canadian School District. *Exceptionality Education International*, *21*(1), 46–60.	1
Menendez, A. L., Payne, L. D., & Mayton, M. R. (2008). The implementation of positive behavioral support in an elementary school: Processes, procedures, and outcomes. *Alberta Journal of Educational Research*, *54*(4), 448–462.	1
Muscott, H. S., Mann, E. L., & LeBrun, M. R. (2008). Positive behavioral interventions and supports in New Hampshire: Effects of large-scale implementation of schoolwide positive behavior support on student discipline and academic achievement. *Journal of Positive Behavior Interventions*, *10*(3), 190–205.	1
Nelson, J. R., Martella, R. M., & Marchand-Martella, N. (2002). Maximizing Student Learning: The Effects of a Comprehensive School-Based Program for Preventing Problem Behaviors. *Journal of Emotional and Behavioral Disorders*, 10(3), 136–148.	1
Sailor, W., Zuna, N., Choi, J. H., Thomas, J., McCart, A., & Roger, B. (2006). Anchoring schoolwide positive behavior support in structural school reform. *Research and Practice for Persons with Severe Disabilities*, *31*(1), 18–30.	1
Schonfeld, D. J., Adams, R. E., Fredstrom, B. K., Weissberg, R. P., Gilman, R., Voyce, C., Speese-Linehan, D. (2015). Cluster-randomized trial demonstrating impact on academic achievement of elementary social-emotional learning. *School Psychology Quarterly*, *30*(3), 406–420.	3

Study	Reason for exclusion
Simonsen, B., Eber, L., Black, A. C., Sugai, G., Lewandowski, H., Sims, B., & Myers, D. (2012). Illinois statewide positive behavioral interventions and supports: Evolution and impact on student outcomes across years. *Journal of Positive Behavior Interventions, 14*(1), 5–16.	1
Wills, H., Kamps, D., Abbott, M., Bannister, H., & Kaufman, J. (2010). Classroom observations and effects of reading interventions for students at risk for emotional and behavioral disorders. *Behavioral Disorders, 35*(2), 103–119.	1
Yeung, A. S., Mooney, M., Barker, K., & Dobia, B. (2009). Does School-Wide Positive Behaviour System Improve Learning in Primary Schools? Some Preliminary Findings. *New Horizons in Education*, *57*(1), 17–32.	1

Appendix B

Weight of Evidence A (WoE A)

Weight of Evidence A (WoE A) provides a judgement of the methodological quality and design employed in each study (Gough, 2007). In order to consistently assess WoE A for each study, an adapted version of the Group based-design coding protocol from Kratochwill's (2003) American Psychological Association Task Force on Evidence-Based Interventions in School Psychology was used. Adaptions were made to the protocol by removing specific components that were not relevant to this systematic literature review and to the studies included. Each study was critically reviewed regarding it's: measurement, comparison, statistical significance, external validity, identifiable intervention components and implementation fidelity. An average (mean) of the scores was calculated to achieve an overall WoE A score.

Table 1 below outlines the measurement criteria. Table 2 includes the Comparison Group Criteria. Table 3 outlines the primary outcome and significance criteria. Table 4 details the external validity criteria. Table 5 defines the implementation fidelity criteria. Table 6 defines the average score ranges. Table 7 summarises overall WoE A scores for each study, and Table 8 outlines the adaptations made to the Kratochwill (2003) coding protocol.

Table 1

Measurement Criteria

Weighting	**Criteria**
Strong evidence (3)	• A reliability coefficient of ≥.85

Weighting	Criteria
	• Collected data using multiple methods • Collected data from multiple sources
Promising evidence (2)	• A reliability coefficient ≥.70 for at least 75% of primary measures • Collected data using multiple methods and/or multiple sources
Weak evidence (1)	• A reliability coefficient of ≥.50 for at least 50% of the primary outcome measures • Collected data uses single method and source
No evidence (0)	• A reliability coefficient of ≤.50 • Collected date from single source and/or data collected using single method.

Table 2

Comparison Group Criteria

Weighting	Criteria
Strong evidence (3)	• At least one type of "active" comparison group must be used • Initial group equivalency must be established (preferably through random assignment of participants) • Evidence that change agents were counterbalanced • Less than 20% attrition.
Promising evidence (2)	• Presence of at least a "no intervention group" Evidence of at least two: • counterbalancing of change agents • group equivalence established • equivalent mortality with low attrition

Weighting	Criteria
Weak evidence (1)	• Presence of a comparison group and at least one: • counterbalancing of change agents • group equivalence established • equivalent mortality with low attrition
No evidence (0)	• No efforts made to ensure group equivalence.

Table 3

Primary Outcome Statistical Significance Criteria

Weighting	Criteria
Strong evidence (3)	• Appropriate statistical analysis must have been conducted, including appropriate units of analysis familywise/experimentwise error rate controlled • A sufficiently large *N* • Must show significant primary outcomes for at least 75% of the total primary outcome measures for each key construct.
Promising evidence (2)	• Appropriate statistical analysis must have been conducted, including appropriate units of analysis familywise/experimentwise error rate controlled • Must show significant primary outcomes for at least 50% to 74% of the total primary outcome measures for each key construct.
Weak evidence (1)	• Appropriate statistical analysis must have been conducted, including appropriate units of analysis familywise/experimentwise error rate controlled • Must show significant primary outcomes for at least 25% to 49% of the total primary outcome measures for each key construct.
No evidence (0)	• None of the above criteria met.

Table 4

External Validity Criteria

Weighting	Criteria
Strong evidence (3)	• Complete and detailed description of the context within which the intervention occurs • Provided evidence of perceived benefits from the intervention for all participant groups.
Promising evidence (2)	• Detailed description of some but not all contextual components • Provided evidence of perceived benefits from the intervention for some participant groups.
Weak evidence (1)	• Provides overview of contextual components but lack details • Provided evidence that participants did not perceive benefits from the intervention.
No evidence (0)	• No description of context • Did not investigate participants' perceptions of benefits.

Table 5

Implementation Fidelity Criteria

Weighting	Criteria
Strong evidence (3)	• Study demonstrates strong evidence of acceptable adherence • Evidence should be measured through at least two of the following: ongoing supervision/consultation, coding sessions, or audio/video tapes, and use of a manual. To be considered a ―manual for a rating of 3, information must have been provided to the implementers using either: written materials involving a detailed account of the exact procedures and the sequence in which they are to be used or a formal training session that includes a detailed account of the exact procedures and the sequence in which they are to be used.

Weighting	Criteria
Promising evidence (2)	• Study must demonstrate evidence of acceptable adherence • Evidence should be measured through at least one of the above criteria and use of a manual. To be considered a ―manual for a rating of 2, information must have been provided to the implementers using either: written materials involving an overview of broad principles and a description of the intervention phases, or a formal or informal training session involving an overview of broad principles and a description of the intervention phases.
Weak evidence (1)	• Study must demonstrate evidence of acceptable adherence measured through at least one of the above criteria or use of a manual.
No evidence (0)	• Nothing done to ensure implementation fidelity or evidence indicates unacceptable adherence.

Table 6

Average (Mean) Score Range for WoE A

Overall Quality	Average Score
High	≥2.5
Medium	1.5 – 2.4
Low	≤1.4

Table 7 *Overall WoE A Scores for Each Study*

Study	Measure	Comparison Group	Statistical significance	External validity Indicators	Identifiable Intervention components	Implementation Fidelity	Overall WoE A
Angus & Nelson (2019)	2	0	3	1	1	2	1.5
Borgen et al. (2019)	0	0	1	1	1	1	0.66
Freeman et al. (2016)	1	1	1	1	1	1	1.3
Houchens et al. (2017)	3	3	3	2	2	2	2.5
Madigan et al. (2016)	2	3	3	2	1	2	2.16
Reno et al. (2017)	3	1	3	1	1	1	1.66
Ryoo et al. (2018)	3	2	3	2	1	2	2.0

Table 8

Amendments Made to the Kratochwill (2003) Coding Protocol

Elimination	**Rationale**
General study characteristics A1 – A5.4	The review has only included quantitative studies. Theoretical basis, validity, nature of research, target group, implementation and interpretation of findings are all discussed in the review.
C7: Coding C8: Interactive process followed	Only necessary for qualitative research
C9: Rival interpretations	This research is only exploring the effectiveness of one intervention on one dependent variable
Key Features for Coding Studies and Rating Level of Evidence/ Support A1: Characteristics of the data collector A2: Characteristics of Participants A3: Sample appropriate to research methods A4: Operationalization. A5: Integration of data from multiple sources	These are not relevant as the participants are whole school settings and their pupils and staff. In this study, data is only generated by the school settings.
B4: Extent of Engagement	As a whole school intervention lead by staff and modelled by senior leaders, this measure if not required. Staff fidelity is measured and included.
B6: Cultural Appropriateness of the Measures	The intervention is universal and considered appropriate for all children and young people.

Elimination	Rationale
Rating for Primary Outcomes Statistically Significant D3, D4, D5, D6	Secondary outcomes are not being measured in this review.
E. Cultural Significance	Not necessary to record for universal school intervention.
F. Educational/Clinical Significance	This is discussed at length during the report
Durability/Generalization of Intervention and Outcomes H1. Follow-up assessment H2. Durability/Generalization over time H3. Durability/Generalization across settings H4. Generalization across persons	Not applicable for longitudinal studies / ongoing whole school approach. Fidelity score is measured over time intervals. Studies are based on implementation in mainstream school settings.

Appendix C

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

Coding Protocol

Date: 29 December 2019

Full Study Reference:

Madigan, K., Cross, R. W., Smolkowski, K., & Strycker, L. A. (2016). Association between schoolwide positive behavioural interventions and supports and academic achievement: a 9-year evaluation. *Educational Research and Evaluation*, *22*(7–8), 402–421.

Intervention Name: Foundations, school wide positive behavioural interventions and supports (PBIS)

Study ID Number: #5

☐ Type of Publication:

☐ Book/Monograph

■ Journal Article

☐ Book Chapter

☐ Other (specify):

1. General Characteristics

A. General Design Characteristics

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

☐ Completely randomized design

☐ Randomized block design (between participants, e.g., matched classrooms)

☐ Randomized block design (within participants)

☐ Randomized hierarchical design (nested treatments

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

☐ Nonrandomized design

☐ Nonrandomized block design (between participants)

☐ Nonrandomized block design (within participants)

☐ Nonrandomized hierarchical design

■ Optional coding for **Quasi-experimental designs**:

The Repeated-Treatment Design (using longitudinal data)

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

☐ Very low (little basis)

☐ Low (guess)

☐ Moderate (weak inference)

☐ High (strong inference)

■ Very high (explicitly stated)

☐ N/A

☐ Unknown/unable to code

B. Data Analysis

	Yes	No
Appropriate unit of analysis	■	☐
Familywise error rate controlled		☐
Sufficiently large *N*		☐

Total size of sample (start of study): 49 schools.

Intervention group sample size:	21 schools (15 elementary, 5 middle, and 1 high). A total of 11,202 pupils.
Control group sample size:	28 comparable schools (21 elementary, 5 middle, and 2 high) A total of 14,857 pupils

C. Type of Program

☐ Universal prevention program

☐ Selective prevention program

☐ Targeted prevention program

■ Intervention/Treatment

☐ Unknown

D. Stage of Program

☐ Model/demonstration programs

☐ Early stage programs

■ Established/institutionalized programs

☐ Unknown

E. Concurrent or Historical Intervention Exposure

■ Current exposure

☐ Prior exposure

☐ Unknown

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

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

A. Measurement

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

■Yes

☐No

☐Unknown/unable to code

Multi-method (at least two assessment methods used)

■ Yes

☐ No

☐ N/A

☐ Unknown/unable to code

Specify: student test scores and school's academic index score.

Multi-source (at least two sources used self-reports, teachers etc.)

■ Yes

☐ No

☐ N/A

☐ Unknown/unable to code

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

☐ Yes validated with specific target group

■ In part, validated for general population only

☐ No

☐ Unknown/unable to code

Measures of key outcomes are linked to the conceptual model.

■ 3 Clear links established between the conceptual model and key outcome indicators

☐ 2 Some, but not all, key outcomes are clearly linked to conceptual model.

☐ 1 Vague reference to links between key outcomes and conceptual model

☐ 0 No evidence that key outcomes are linked to conceptual model.

Overall Rating for measurement (select: 0,1, 2 or 3)	3	2	1	0
	☐	■	☐	☐

B. Comparison Group

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

☐ Typical intervention (typical intervention for that setting, without additions that make up the intervention being evaluated)

☐ Attention placebo

☐ Intervention element placebo

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

Specify: traditional approaches to behaviour management and discipline.

B2 Overall confidence of judgment on type of comparison group

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

B3 Counterbalancing of change agent (participants who receive intervention from a single therapist/teacher etc were counter-balanced across intervention)

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

B4 Group equivalence established (select one of the following)

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

B5 Equivalent mortality

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

Overall Rating for comparison group (select: 0,1, 2 or 3) 3 ■ 2 ☐ 1 ☐ 0 ☐

C. Primary/Secondary Outcomes Are Statistically Significant

C1. Evidence of appropriate statistical analysis for **primary outcomes**

■ Appropriate unit of analysis (rate from previous code)

■ Familywise/experiment wise error rate controlled when applicable (rate from previous code)

■ Sufficiently large *N* (rate from previous code)

C2. Percentage of **primary outcomes** that are significant (select one of the following)

■ Significant primary outcomes for at least 75% of the total primary outcome measures for each key construct

☐ Significant primary outcomes for between 50% and 74% of the total primary outcome measures for each key construct

☐ Significant primary outcomes for between 25% and 49% of the total primary outcome measures for any key construct.

Rating for Primary Outcomes Statistically Significant 3 ■ 2 ☐ 1 ☐ 0 ☐

OVERALL Rating for Primary/Secondary Outcomes 3 ■ 2 ☐ 1 ☐ 0 ☐

External Validity Indicators

Sampling Procedures

Sampling procedures described in detail

■ Yes

☐ No

Rationale for sample selection specified

■ Yes

Specify: County Superintendent's Office identified 26 schools (16 elementary, 8 middle, and 2 high) to participate in either Treatment Cohort I or Treatment Cohort II, based on involvement with the intervention initiative. Only schools with moderate to high fidelity of implementation were included in the present study.

☐ No

Rationale for sample size specified

■ Yes

Specify: As above, schools from the state county were only eliminated from intervention group for non-involvement with the programme or for low fidelity scores in program implementation.

☐ No

Evidence provided that sample represents target population

■ Yes

☐ No

Rating for Sampling

3	2	1	0
■	☐	☐	☐

Adequately reported characteristics of participants/sample. Adequate level of detail in description of participants

■ Yes

☐ No

Details are provided regarding variables that:

Have differential relevance for intended outcomes ☐ Yes ■ No

Specify: N/A.

Have relevance to the inclusion criteria ■ Yes ☐ No

Specify: variables relate directly to aim of study: intervention impact school academic index score, which is based on student test scores. However, the variable thought to result in improved academic achievement were not measured.

Transferability of the intervention

☐ 3 Complete and detailed description of the context within which the intervention occurs

☐ 2 Detailed description of some but not all contextual components

■ 1 Provides overview of contextual components but lack details

☐ 0 No description of context

Participant perceptions of benefits of intervention (treatment group)

☐ 3 Provided evidence of perceived benefits from the intervention for all participant groups

☐ 2 Provided evidence of perceived benefits from the intervention for some participant groups

☐ 1 Provided evidence that participants did not perceive benefits from the intervention.

■ 0 Did not investigate participants' perceptions of benefits.

OVERALL Rating for External Validity

3	2	1	0
☐	■	☐	☐

Identifiable Intervention Components

Overall Rating for Identifiable Components:

3	2	1	0
☐	☐	■	☐

Design allows for analysis of identifiable components ☐ Yes ■ No

Total number of components: 1 / N

Number of components linked to primary outcomes: 1 / N

Clear documentation of essential components ☐ Yes ■ No

Specify: The study discusses the theoretical underpinnings of the intervention. It explains that schools must engage with specific training and remain actively involved with the programme to contribute data and ongoing trainings. The paper signposts the reader to a website for technical assistance. However, detailed components of the training are not listed in the paper.

Procedures for adapting the intervention are described in detail ☐ Yes ■ No

Specify:

Contextual features of the intervention are documented ■ Yes ☐ No

Specify: Introduction of the paper gives a clear description of the intervention. The paper specifies that in order to deliver the intervention: 'the school's leadership team attends the 15 to 20 days of training spread over multiple years; the leadership team then trains the remaining school personnel', and fidelity to the intervention in monitored.

OVERALL Rating of Identifiable Intervention Components

3	2	1	0
☐	☐	■	☐

D. Implementation Fidelity

Evidence of Acceptable Adherence

■ Ongoing supervision/consultation

☐ Coding intervention sessions/lessons or procedures

☐ Audio/video tape implementation

- ☐ Entire intervention
- ☐ Part of intervention

Manualization (select all that apply)

☐ Written material involving a detailed account of the exact procedure and the sequence they are to be used.

■ Formal training session that includes a detailed account of the exact procedures and the sequence in which they are to be used.

■ Written material involving an overview of broad principles and a description of the intervention phases.

■ Formal or informal training session involving an overview of broad principles and a description of the intervention phases.

Adaptation procedures are specified (select one)

☐ Yes ■ No ☐ unknown

Rating for Implementation Fidelity

3	2	1	0
☐	■	☐	☐

Summary of Evidence

Indicator	Overall evidence rating 0-3	Description of evidence Strong Promising Weak No/limited evidence Or Descriptive ratings
General Characteristics		
General design characteristics		Promising: Quasi-experimental study using Longitudinal data and large N, with good control group. However, the design did not allow for causal attribution.
Data analysis		Strong: Multivariate analysis of covariance (MANCOVA), with two school years of baseline measurements prior to intervention measurement. .
Type of program		Whole school intervention
Stage of program		Established
Concurrent/historical Intervention Exposure		Current exposure
Key Features		
Measurement	2	Promising evidence. Thorough baseline measurements and comparison with a control group. Limitation was lack of measurement for causality
Comparison	3	Strong evidence: good use of a three tiered matching procedure
Primary outcomes are statistically significant	3	Strong evidence: $P < .001$ with large effect size

External validity indicators	2	Promising evidence: based on sampling, ecological validity and experimental design using longitudinal data. Lack of detail regarding contextual components
Identifiable Intervention components	1	Description and signposting only. Not possible to measure specific components of the intervention but intervention components could have been more identifiable by exploring self-report.
Implementation Fidelity	2	Only schools that are able to demonstrate fidelity consistently were included.

Appendix D

Weight of Evidence B (WoE B): provides a review specific judgement of the relevance of the methodology in relation to the review question (Gough, 2007). Typologies, are considered to be helpful in organising and appraising evidence and may be considered as more helpful than hierarchies (Petticrew and Roberts, 2003). Petticrew and Roberts (2003) typology highlights that Randomised Control Trials are best suited to address research questions based on 'effectiveness, followed by quasi-experimental studies and cohort studies. All of the studies in this review fall under the category of quasi-experimental studies and the WOE B criteria has been selected to reflect this. Table 1 below details the criteria that each of the included studies must meet in order to receive the specified rating. To achieve a specified rating, four of the five criteria must be met by a study.

Table 1

WoE B Weighting Criteria

Weighting	**Criteria**
Strong/High (3)	• Equivalent group design is used • Three or more pupil cohorts or whole schools are included for treatment and control • Pretest-Postest measures include three or more academic subjects / learning areas for three or more pupil cohorts • Time series measurements used • Measures used to test effectiveness are very clearly outlined and include reliability and validity details.
Promising/ Medium (2)	• Non-equivalent group design with steps taken to ensure multiple similarities between treatment and control group • Two pupil cohorts or whole schools are included for treatment and control • Pretest-Postest measures include two academic subjects

Weighting	Criteria
	• Time series measurements used • Measures used to test effectiveness are reported and include reliability or validity details.
Weak /Low (1)	• No control group or non-equivalent group and no efforts made to ensure similarities between treatment and control group • One pupil cohort or school is included for treatment and control • Pretest-Postest measures for one academic subject • No use of time series measurements • Measures used to test effectiveness are reported.

Table 2

Weight of Evidence B

Study	Overall WoE B
Angus & Nelson (2019)	1
Borgen et al. (2019)	2
Freeman et al. (2016)	2
Houchens et al. (2017)	3
Madigan et al. (2016)	3
Reno et al. (2017)	2
Ryoo et al. (2018)	2

Appendix E

Weight of Evidence C (WoE C): provides a judgement of the topic relevance in relation to the review question (Gough, 2007). The criteria for WoE C is listed in table 1, below. To achieve a specified rating, five of the six criteria must be met by a study.

Table 1

Criteria and Weighting Table

Weighting	Criteria
Strong/High (3)	• The study specifically measures the intervention effectiveness at improving academic performance • Intervention fidelity is measured by an external PBS professional using an assessment measure published specifically for PBS, i.e. the Tiered Fidelity Inventory (TFI), Facility-Wide Tiered Fidelity Inventory (FW-TFI), Benchmarks of Quality (BoQ), School-wide Evaluation Tool (SET), Commitment and Implementation sections of the PBS Framework. • The measurement school has ongoing PBS training and access to a coach, guide or mentor. Either training or coaching sessions take place a minimum of once every school term or semester. • Academic achievement measure is based on a criterion-reference test, such as SATs, GCSE, or state equivalent in the US, such as high school diploma. The measure includes a minimum of three subjects (English, maths and science). • Children/Young People of whom academic achievement is measured are from a minimum of three year groups, cohorts or schools. • The study includes data regarding the school demographic (cultural backgrounds and Social Economic Status).
Promising/ Medium (2)	• The study specifically measures the intervention effectiveness at improving academic performance • Intervention fidelity is measured by a member of school staff using an assessment measure published specifically for PBS, i.e. the Tiered Fidelity Inventory (TFI), Facility-Wide Tiered Fidelity Inventory (FW-

Weighting	Criteria
	TFI), Benchmarks of Quality (BoQ), School-wide Evaluation Tool (SET) • The measurement school has ongoing PBS training and access to a coach, guide or mentor. Either training or coaching sessions take place a minimum of once every school term or semester. • Academic achievement measure is based on a criterion-reference test, such as SATs, GCSE, or state equivalent in the US, such as high school diploma. The measure includes a minimum of two subjects (English and maths). • Children/Young People of whom academic achievement is measured are from a minimum of two year groups, cohorts or schools. • The study includes data regarding the school demographic (cultural backgrounds and Social Economic Status).
Weak /Low (1)	• The study specifically measures the intervention effectiveness at improving academic performance • Intervention fidelity measure is reported but it is unclear what tool has been used, or a generalised checklist has been used. The measure is based on school self-assessment/report. • The measurement school has a trained/qualified member of staff working within the school as an additional employment duty and provides coaching, mentoring or guidance to other staff. • Academic achievement measure is based on a curriculum based measure or a specific assessment e.g. a literacy or numeracy assessment. • Children/Young People of whom academic achievement is measured are from a minimum of two year group, cohorts or schools. • The study includes data regarding the school demographic (cultural backgrounds and Social Economic Status).
No evidence (0)	• There is insufficient evidence that the study meets any of the criteria above.

Table 2

Rationale for WoE C Criteria

Criteria	Rationale
Intervention effectiveness	The study must focus on academic achievement / attainment in order to be aligned with the review question.
Fidelity measure	Intervention fidelity must be measured to ensure construct validity of the study. The Tiered Fidelity Inventory (TFI), Facility-Wide Tiered Fidelity Inventory (FW-TFI), Benchmarks of Quality (BoQ) and School-wide Evaluation Tool (SET) are widely recognised as standard measures for positive behaviour support. Bias is reduced and reliability increased when the measure is carried out by an external and appropriately trained professional.
Continuing Professional Development	To ensure that intervention fidelity is longitudinal and that the schools use of the intervention is reliable, there should be ongoing coaching/training/mentoring of positive behaviour support.
Academic achievement measure	Core subjects offer consequential validity and predictive validity.
Reliability of academic progress	Increasing the measure over time increases the chance of measuring the impact of intervention, and reduces the likelihood of a result occurring by chance.
Demographic data	Demographic data is important for the determination of whether the individuals in each study are a representative sample for generalization purposes.

Table 3
Weight of Evidence C

Study	Overall WoE C
Angus & Nelson (2019)	3
Borgen et al. (2019)	2
Freeman et al. (2016)	1
Houchens et al. (2017)	2
Madigan et al. (2016)	3
Reno et al. (2017)	1
Ryoo et al. (2018)	3

Appendix F

Table 1

Summary of weight of evidence

Study	WoE A	WoE B	WoE C	WoE D
Angus & Nelson (2019)	1.5 (Medium)	1 (Low)	3 (High)	1.83 (Medium)
Borgen et al. (2019)	0.66 (Low)	2 (Medium)	2 (Medium)	1.55 (Medium)
Freeman et al. (2016)	1.3 (Low)	2 (Medium)	1 (Low)	1.43 (Low)
Houchens et al. (2017)	2.5 (High)	3 (High)	2 (Medium)	2.5 (High)
Madigan et al. (2016)	2.16 (Medium)	3 (High)	3 (High)	2.72 (High)
Reno et al. (2017)	1.66 (Medium)	2 (Medium)	1 (Low)	1.55 (Medium)
(Ryoo et al. (2018)	2.0 (Medium)	2 (Medium)	3 (High)	2.3 (Medium)

*WoE D rating descriptors: ‘High’ ≥2.5, ‘Medium’ 1.5 –2.4, ‘Low’ ≤1.4