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

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

How effective is the Good Behaviour Game at improving social and behavioural outcomes for primary aged pupils?

Summary

The Good Behaviour Game (GBG) is an interdependent group contingency intervention that is delivered at the whole-class level. The intervention has been originally used to reduce disruptive behaviour in schools but has recently been adapted to target a variety of other behaviours such as encouraging positive social behaviour (Bowman-Perrot et al., 2015). A systematic literature review was conducted to determine the effectiveness of the GBG on improving social and behavioural outcomes in primary aged pupils. Eight studies were selected based on the inclusion criteria and were evaluated in line with Gough's (2007) Weight of Evidence Framework and the coding protocol devised by Gersten et al. (2005) to enable a critical review of the evidence. The review found insufficient evidence to conclude that the GBG is an effective intervention for improving the social and behavioural outcomes for primary aged pupils. This is demonstrated by inconsistent significant results with negligible effect sizes. Limitations of the review are highlighted, along with recommendations for future research.

Introduction

The Good Behaviour Game

The Good Behaviour Game (GBG) is a universal behaviour management intervention that employs an interdependent group contingency to increase desirable behaviours in the classroom (Bowman-Perrot et al., 2015). This form of group contingency involves a reward system contingent upon the collective group performance (Tingstrom et al., 2006). The GBG involves playing a game during a lesson, where the class work together in teams to self-regulate their own behaviour as well as their peers so that they can access an individual or group reward (Coombes et al., 2016). The GBG follows four core principles, involving creating and adhering to classroom rules, being a part of a team, behaviour monitoring, and access to positive reinforcements (Chan et al., 2012). Initially, the class is divided into teams with each team balanced by behaviour, gender and academic ability to ensure equivalent groups. The class teacher then outlines the rules of the game which involve a description of desired classroom behaviour, such as '*work quietly'*. For every rule that is broken, the class teacher will give a point to that team. The team that scores the least amount of points by the end of the lesson will win the game (Coombes et al., 2016).

The GBG incorporates a natural progression in terms of how it is implemented. To begin with, games are played three times a week, for 10 minutes across the academic year, gradually increasing to games being played daily for the

duration of a lesson (Humphrey et al., 2018). Additionally, rewards will be given to the winning team using tangible reinforcements, such as stickers, which are then replaced by less tangible rewards and extrinsic rewards over time, including golden time and verbal praise (Coombes et al., 2016). The aim of this, is to increase pupils' motivation, interest and enable behaviours to be generalised across other activities (Humphrey et al., 2018).

Originally developed in the United States (US) by Barrish et al. (1969), the GBG has been implemented in elementary schools to reduce disruptive behaviour such as, out-of-seat and talking-out behaviour. Since its conception, the GBG has been implemented to reduce a variety of problem behaviours including, oppositional behaviour (Leflot et al., 2010) aggression and bullying (Kellam et al., 2008). More recently, the GBG has been evaluated to assess its impact on promoting positive social behaviours with peers (Sewell, 2020), based on the assumption that peer acceptance is likely to increase when pupils try to help and support each other to display desirable behaviour (Breeman et al., 2016). The GBG has also been subject to various modifications, such as providing points for displaying positive behaviours, and has been implemented in a variety of setting such as, cafeterias (Bowman-Perrott et al., 2016) and in afterschool clubs (Smith et al., 2018). In a review of single-case experimental designs, Bowman-Perrot et al. (2016) reported that modified versions of the GBG are equally effective, as long as adherence to the interdependent group contingency strategy and the four key principles, outlined above, are maintained.

The advantages of the GBG is that it is a behaviour management strategy that is easy to implement and can be played alongside the curriculum, without

compromising valuable teaching time (Chan et al., 2012). Additionally, it offers teachers flexibility in terms of the behaviours they choose to target, when games are played, as well as the type of reinforcements that work best for the class (Bowman-Perrott et al., 2016).

Psychological Basis of the GBG

The GBG is underpinned by three theories for understanding human behaviour. Firstly, principles from Behaviourism (Skinner, 1945) suggest that positive behaviours are likely to be reproduced if followed by a positive reinforcement. Similarly, undesirable behaviours are less likely to reoccur if the behaviour is followed by a negative reinforcement such as, a punishment or sanction. The use of a reward system in the GBG given to the team who displays the least amount of undesirable behaviour aims to ensure positive behaviours are reproduced and maintained. Given that the GBG uses a group contingency, the intervention is also underpinned by Bandura's (1977) Social Learning Theory, in that pupils can learn appropriate behaviours from model peers. Lastly, the GBG draws upon the Life Course/Social Field theory (Kellam et al., 2011) which proposes the idea that social tasks demand, such as complying with rules, paying attention and interacting appropriately, are associated with successful adaptations in other social situations. Within this context, the GBG allows teachers to directly teach socially acceptable behaviours to help prepare pupils for social contexts beyond the classroom (Humphrey et al., 2018).

Rationale and Relevance

The presence of low-level disruptive behaviour in the classroom, such as calling out and off-task behaviour, can have a detrimental impact on learning, participation in lesson's and overall academic outcomes (Ofsted, 2014). It was

reported by the Office for Standards in Education (2014) that approximately one hour of learning is lost daily due to disruptive behaviour in the classroom. In the United Kingdom (UK), the rate of exclusions has increased from 2018 to 2019, with persistent disruptive behaviour accounting for 31% of fixed-term exclusions and 35% of permanent exclusions across state-funded schools (DfE, 2021). Not only can disruptive behaviour impact upon pupil outcomes, it can also interfere with the learning of others and has led to the reduction of teachers in the profession (Ofsted, 2014).

Additionally, pupils displaying disruptive behaviour also tend to demonstrate poor relationships with others and low social competence (Hukkelberg et al., 2019). Low social competence has been associated with poorer emotional wellbeing in childhood and can lead to disruptive behaviour being displayed in school. Furthermore, children who receive social skills support are less likely to develop psychiatric disorders in adulthood (Sewell, 2019; Coombes et al., 2016).

As well as a rise in exclusion rates, there has been increasing concerns in regards to children's mental health, with more children experiencing social, emotional and behavioural difficulties in schools (DHSC & DfE, 2017). This emphasises the need for schools to deliver evidence-based interventions to support children at the individual, group and whole-school level so that risks to mental health and school exclusions are prevented.

In the UK, Educational Psychologist's (EPs) support schools with the implementation of evidence-based interventions (HCPC, 2016), and play a key role in facilitating change in social, emotional and mental health outcomes at the

individual, group and universal level (DHSC & DfE, 2017; DfE & DoH, 2015). The GBG provides a universal intervention that can be used as a behavioural management system to decrease disruptive behaviour (Barrish et al., 1969) and increase positive social behaviour (Sewell, 2020). Therefore, the purpose of this will review will be of relevance to EP practice as the findings will provide an overview of the evidence-base on the impact of the GBG game on social and behavioural outcomes, which can be offered to schools to prevent the rise in social and behavioural difficulties present in school.

While the GBG has undergone extensive research and has been evaluated in numerous literature reviews and meta-analyses (Flower et al., 2014; Bowman-Perrott et al., 2016; Tingstrom et al., 2006), it has been concluded that the methodological rigour of previous findings were not enough to meet standards for evidence-base practice (Bowman-Perrott et al., 2016). Thus, highlighting the need for a review of research using more robust methodology that compares the effects of the GBG with a control group. Additionally, the focus of previous reviews has been on disruptive behaviour, with little known about the overall effects of the GBG on promoting positive social behaviour (Coombes et al., 2016; Sewell, 2020). This demonstrates the need to conduct a review to explore the impact of the GBG on social and disruptive behaviour.

Review Question

How effective is the GBG at improving social and behavioural outcomes for primary aged pupils?

Critical Review of the Evidence Base

Literature Search

A systematic literature search was carried out on 5th January 2021 using the following three databases: PsycINFO, Web of Science and Educational Resources Information Center (ERIC). Table 1 outlines the search terms used in the literature search.

Table 1.

Search Terms for Literature Search

Intervention		Context		Participants
"Good Behaviour Game" OR GBG	AND	School	AND	Pupil OR Student OR Children

Note: The use of 'AND' combines search term so that results include both terms. Search terms separated by 'OR' ensures that results consider alternative terms of the same concept (e.g. pupil OR student). Quotation marks yield results for the exact phrase of concepts (e.g. "Good Behaviour Game").

Article Screening

Figure 1 represents a flow diagram of the literature search and article screening process. Where possible, search results were filtered to studies that had been peer reviewed (PsycINFO and ERIC) and published between 2015 to 2021 (PsycINFO and Web of Science). The rationale for this is outlined in the inclusion and exclusion criteria in Table 2. This yielded a total of 276 studies. After the removal of duplicates, the titles and abstracts of 262 studies were screened against the inclusion and exclusion criteria to determine eligibility for review. This led to the exclusion of 243 studies, with full-text screening carried out on the remaining 19 studies. Studies excluded at the full-text screening

stage are presented in Appendix A with the reasons for exclusion. A total of 8 studies met the eligibility criteria and were selected for in depth review (see Table 3 for included studies).

Figure 1:

Flow Chart of the Systematic Literature Search Process

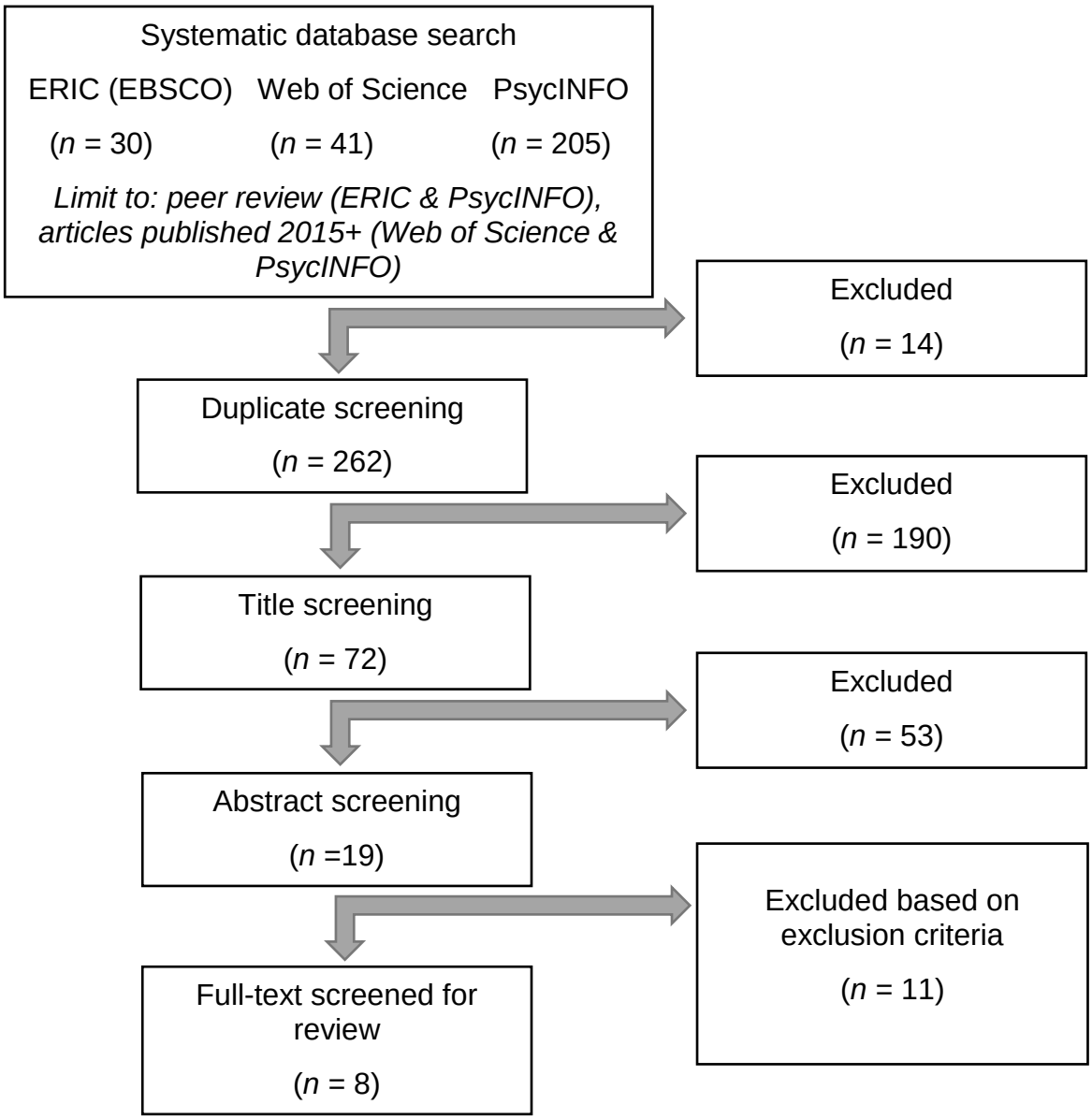

Table 2.

Inclusion and Exclusion for Study Screening

Study Feature	Inclusion Criteria	Exclusion Criteria	Rationale
1. Intervention	Good Behaviour Game or adapted version of the GBG	Good Behaviour Game is not included in the study	The research question aims to examine the effectiveness of the GBG intervention
2. Participants	Study includes pupils of primary school age (5 to 11 years)	Participants not of primary school age	Research questions aims to address the intervention effect on primary aged pupils which has been the main focus of previous GBG research (Bowman-Perrott et al., 2016)
3. Publication type	Peer reviewed study	Non-peer reviewed studies	To ensure study has been rigorously reviewed
4. Research design	Experimental design with comparison group	Experimental studies without a comparison group, qualitative research, case study designs	To date, there has been no review conducted on studies that use experimental designs with a comparison group. Comparison studies will help to explore whether the outcomes are associated with the intervention or not.
5. Outcomes	Pre- and post-intervention measures assessing social and behavioural outcomes	Study presents no pre- and post- intervention measures for social and behavioural outcomes	To review the effects of the intervention on pupils' social and behavioural outcomes
6. Language	Study published in English	Study published in a language other than English	This review has not got access to software to translate articles
7. Date of publication	Study published in 2015 onwards	Studies published prior to 2015	This review aims to examine recent findings on the GBG. Studies published in the past six years was considered appropriate by author.

Table 3.

Full Reference of Studies included in the Review

Studies included review
Ashworth, E., Humphrey, N., & Hennessey, A. (2020). Game Over? No Main or Subgroup Effects of the Good Behavior Game in a Randomized Trial in English Primary Schools. *Journal of Research on Educational Effectiveness, 13*(2), 298–321.
Bradshaw, C. P., Shukla, K. D., Pas, E. T., Berg, J. K., & Ialongo, N. S. (2020). Using Complier Average Causal Effect Estimation to Examine Pupil Outcomes of the PAX Good Behavior Game When Integrated with the PATHS Curriculum. *Administration and Policy in Mental Health and Mental Health Services Research, 47*(6), 972–986.
Breeman, L. D., van Lier, P. A. C., Wubbels, T., Verhulst, F. C., van der Ende, J., Maras, A., Struiksma, A. J. C., Hopman, J. A. B., & Tick, N. T. (2016). Effects of the Good Behavior Game on the Behavioral, Emotional, and Social Problems of Children with Psychiatric Disorders in Special Education Settings. *Journal of Positive Behavior Interventions, 18*(3), 156–167.
Hart, S. R., Celene, D., Embry, D. D., Becker, K., Lawson, A., & Ialongo, N. (2021). The Effects of Two Elementary School-Based Universal Preventive Interventions on Special Education Pupils' Socioemotional Outcomes: RASE. *Remedial and Special Education, 42*(1), 31–43.
Ialongo, N. S., Domitrovich, C., Embry, D., Greenberg, M., Lawson, A., Becker, K. D., & Bradshaw, C. (2019). A Randomized Controlled Trial of the Combination of Two School-Based Universal Preventive Interventions. *Developmental Psychology, 55*(6), 1313–1325.
Smith, E. P., Osgood, D. W., Oh, Y., & Caldwell, L. C. (2018). Promoting Afterschool Quality and Positive Youth Development: Cluster Randomized Trial of the Pax Good Behavior Game. *Prevention Science, 19*(2), 159–173.
Spilt, J. L., Leflot, G., Onghena, P., & Colpin, H. (2016). Use of Praise and Reprimands as Critical Ingredients of Teacher Behavior Management: Effects on Children's Development in the Context of a Teacher-Mediated Classroom Intervention. *Prevention Science, 17*(6), 732–742.
Streimann, K., Selart, A., Trummal, A., Karin, S., Anne, S., & Aire, T. (2020). Effectiveness of a Universal, Classroom-Based Preventive

Intervention (PAX GBG) in Estonia: a Cluster-Randomized Controlled Trial. *Prevention Science, 21*(2), 234–244.

Mapping the Field

A description of the eight studies included in this systematic literature review is presented in Table 4.

Table 4.

Mapping the Field

Author	Location	Sample & Participant Characteristics	Study Type & Control Group	Measures	Outcomes
Ashworth et al. (2020)	United Kingdom	*N* = 3084 pupils (77 schools) Primary pupils aged between 6-7 years (male: 52.6%, female: 47.4%) Male: 52.6%, Female: 47.4	Cluster Randomised Control Trial Intervention: The GBG Control: Usual provision (schools usual behaviour management system)	<u>Teacher rated:</u> Teacher Observation of Classroom Adaptation (TOCA-C)	No significant main effect of GBG on disruptive behaviour and prosocial behaviour compared to control group.
Bradshaw et al (2020)	United States	*N* = 1526 (27 schools) Elementary pupils across Kindergarten to 5th grade identified displaying disruptive and aggressive behaviour Male: 50.81%, Female: 49.19	Cluster Randomised Control Trial Intervention: PAX GBG Control: Usual provision (schools usual behaviour management system)	<u>Teacher rated:</u> Teacher Observation of Classroom Adaptation-Revised (TOCA-R)	No significant main effects were found for pupils in the GBG condition for hyperactivity, aggressive-disruptive problems, social competence and peer relations.
Breeman et al. (2016)	Netherlands	*N* = 389 (11 schools)	Cluster Randomised Control Trial	<u>Teacher rated:</u>	There was a significant main effect of the intervention on teacher

		Pupils aged 5 - 13 years (M = 10.08 years) attending a special primary education. All pupils had been referred to the setting for significant social, emotional and behavioural problems (e.g. Autism: 44.5%, Attention Deficit Hyperactivity Disorder: 38.8%, Conduct Disorder: 28.0%) Male: 87%, Female: 13%	Intervention: The GBG Control group: Usual provision (schools usual behaviour management system)	Problem Behaviour at School Interview (PBSI) <u>Child rated:</u> Children's Social Preference Procedure	reported behavioural problems. This was demonstrated by a reduction in behavioural problem scores at post-test. No differences were found on pupil-reported social preference.
Hart et al. (2021)	United States	*N* = 650 (27 schools) Elementary pupils receiving special education across K-5 grade Male: 65.2%, Female: 34.8%	Cluster Randomised Control Trial Intervention: PAX GBG Control: *no description of control group given*	<u>Researcher:</u> Independent observation of pupil behaviour <u>Teacher rated:</u> Teacher Observation of Classroom Adaptation-Revised (TOCA-R)	Pupils in the intervention group did not significantly differ from the comparison group in terms of authority acceptance, total problem behaviours and social competence.

				The Social Health Profile Social Competence Scale	
Ialongo et al. (2019)	United States	*N* = 5611 (9 schools) Elementary pupil across Kindergarten to 5th grade Male: 50.81%, Female: 49.19%	Cluster Randomised Control Trial Intervention: PAX GBG Control: Usual provision (schools usual behaviour management system)	<u>Researcher:</u> Independent observation of pupil behaviour <u>Teacher rated:</u> Teacher Observation of Classroom Adaptation-Revised (TOCA-R) The Social Health Profile Social Competence Scale	Pupils in the intervention group demonstrated a reduction in problem behaviour at post-test compared to pupils in the control group. No significant differences were found for social competence.
Smith et al. (2018)	United States	*N* = 811 (76 afterschool programs) Elementary pupils across grades 2-5 (aged 5- 13 years) attending afterschool programs Male: 49.90%, Female: 50.1%	Cluster Randomised Control Trial Intervention: PAX GBG Control: Usual provision (usual behaviour management system)	<u>Child rated:</u> Strengths and Difficulties Questionnaire (SDQ)	Pupils in the intervention group had significantly higher levels of pupil-reported prosocial behaviours at post-test than the control group. No differences were found between control and intervention group for total problem behaviour and conduct problems.

Split et al. (2016)	Belgium	*N* = 570 (15 schools) Elementary pupils with a mean age of 7 years and 5 months Males: 49.5%	Cluster Randomised Control Trial Intervention: The GBG Control: *no description of control group given*	<u>Teacher rated:</u> The Problem Behaviour at School Interview-revised (PBSI) The Asocial Behaviour subscale of Child Behaviour Scale <u>Child rated:</u> Peer nomination assessment	Pupils reported a significant reduction in scores on oppositional behaviour in the intervention group compared to the control group. No significant differences were reported for other behavioural outcomes.
Streimann et al. (2020)	Estonia	*N* = 708 (42 schools) Elementary pupils (aged 7 to 8 years) Males: 49.9%	Cluster Randomised Control Trial Intervention: PAX GBG Control: Wait-list control	<u>Teacher rated:</u> Strengths and Difficulties Questionnaire (SDQ) <u>Parent rated:</u> Strengths and Difficulties Questionnaire (SDQ)	Pupils in the intervention group demonstrated a reduction in conduct problem behaviour in the intervention group at post-test. No differences were found between the intervention and control group for prosocial behaviours.

Weight of Evidence

Gough's (2007) Weight of Evidence (WoE) Framework was employed to evaluate the extent to which the eight included studies contribute to answering the review question. The framework consists of making weighted judgements across three dimensions. The first involves an appraisal of each study based on the quality of the methodology, referred to as Weight of Evidence A (WoE A; Gough, 2007). This required the use of the coding protocol by Gersten et al. (2005) which is suitable for evaluating studies that used experimental group designs. See Appendix B for an example of the coding protocol. Secondly, studies were then evaluated based on the methodological relevance in answering the review question (WoE B; Gough, 2007). A judgement for WoE B was made in line with Petticrew and Roberts' (2003) typology of evidence. Each study was then appraised on how relevant the topic of the study was in answering this review question (WoE C) based on criteria determined by the author. See Appendix C and D for judgement criteria for WoE B and C. Scores were given for each dimension (WoE A, B and C) and were then averaged to produce an overall judgement rating (WoE D; see Table 5).

Table 5.

Weight of Evidence Ratings

Studies	WoE A	WoE B	WoE C	WoE D
Ashworth et al. (2020)	3 (High)	3 (High)	2.75 (High)	2.92 (High)
Bradshaw et al (2020)	3 (High)	3 (High)	2.25 (High)	2.75 (Medium)
Breeman et al. (2016)	1 (Low)	3 (High)	2.5 (High)	2.17 (Low)
Hart et al. (2021)	1 (Low)	3 (High)	2 (Medium)	2.00 (Low)
Ialongo et al. (2019)	3 (High)	3 (High)	2.25 (High)	2.75 (Medium)
Smith et al. (2018)	3 (High)	3 (High)	1.75 (Medium)	2.58 (Medium)
Split et al. (2016)	3 (High)	3 (High)	2.25 (High)	2.75 (Medium)
Streimann et al. (2020)	1 (Low)	3 (High)	2.25 (High)	2.08 (Low)

Note. For WoE D a score of 2-2.5 = low, 2.6-2.7 = medium, 2.8.-3 = high

Participants

The combined total of 13,349 participants took part in the reviewed studies, ranging from a sample of 389 to 3084 across studies. In line with the inclusion criteria, all studies included participants of primary school age and were randomly allocated to experimental or comparison groups at the school or classroom level. To limit the effects on selection bias to internal validity, six studies reported either using matched pairs or stratification methods to ensure equivalency between settings and conditions. Two studies, did not described how randomised allocation occurred to ensure equivalent groups. However, demographic profiles of participants involved in the study were reported (Bradshaw et al., 2020; Split et al., 2016). The inclusion of male and female participants was fairly equivalent across studies, apart from one study where

the proportion of males make up 87% of the sample size (Breeman et al., 2016).

Among the studies, three recruited participants identified as ‘high-risk’ based on evaluated aggressive-disruptive scores at baseline (Bradshaw et al., 2020), those receiving special education (Hart et al., 2021) or diagnosed with a psychiatric disorder associated with social, emotional and mental health needs (Breeman et al., 2016; see Table 4 for outline of diagnoses). Previous findings report that the benefits of the GBG are most salient for pupils who display higher levels of social, emotional and behavioural difficulties (Bowman-Perrot et al., 2016). With this in mind, it is hypothesised that these studies are more likely to produce positive outcomes, thus were given higher WoE C ratings on participant characteristics. The remaining five studies did not report descriptions of difficulties presented at the class or individual level. Therefore, received lower WoE C ratings (see Appendix D, Table 1 & 2).

Setting

One of the studies was conducted in a primary school in the UK (Ashworth et al., 2020). This allows for findings to be generalised to the UK, so received a higher WoE C rating for setting (see Appendix D, Table 1 & 2). Six studies received a medium rating for being conducted in elementary school settings across the US (Bradshaw et al., 2020; Hart et al., 2021), the Netherlands (Breeman et al., 2016; Ialongo et al., 2019), Belgium (Split et al., 2016), and Estonia (Streimann et al.,2020). These settings are Organisation for Economic Co-operation and Development (OECD) countries where education systems are fairly similar, which led to medium ratings. Smith et al. (2018) received a lower

WoE C rating as the intervention was conducted across afterschool programs in the US, thus lacking ecological validity to school settings.

Research Design

All eight studies used Cluster Randomised Control Trials (RCT), identified as a high-quality research design suitable for examining the effectiveness of interventions (Petticrew & Roberts, 2003). Additionally, all studies met criteria for including pre- and post-measures with a comparison group, leading to high WoE B ratings (see Appendix C, Table 1, 2 & 3). Six of the studies included their usual provision which was the behaviour management system currently being used in that setting. Three of the six studies that included usual provision as a comparison group also compared the GBG with an integrated version of the programme. This involved the GBG and the Promoting Alternative Thinking Strategies (PATHS) curriculum which is a socio-emotional literacy programme (Bradshaw et al., 2020; Hart et al., 2021; Ialongo et al., 2019). For the purpose of this review, the GBG was not compared against the integrated version as they are similar in design, thus will not help to answer the review question. Streimann et al. (2020) included a wait-list control which meant the intervention can be compared against the settings usual behaviour management system, whilst also allowing children to benefit from the intervention at a later stage. Comparison groups were not described in two of the eight studies (Hart et al., 2021; Split et al., 2016). Therefore, little is known about the differences between outcomes of the GBG and the comparison group.

Intervention

The origins of the intervention varied across the studies. Only one of the eight studies (Ashworth et al., 2020) implemented the original version of the GBG

without any adaptations or modifications, yielding a higher rating for WoE C (see Appendix D, Table 2). Rather than assigning points to teams displaying disruptive behaviour, Breeman et al. (2016) and Split et al. (2020) used an adapted version of the GBG whereby teams were presented with cards at the start of the game. The removal of cards was contingent upon disruptive behaviour without much attention being paid to said behaviour. The teams with the most cards by the end of the game received the reward.

Hart et al. (2021), Ialongo et al. (2019), Smith et al. (2018) and Streimann et al. (2020) all used the PAX version of the GBG. This version involves the implementation of 'PAX kernels' both during and outside of the game. These are a set of evidence-based practices that influence behavioural and psychological processes. Examples of 'PAX kernels' involve providing a timer to support children to focus attention and engagement on a task, a cue demonstrating expected level of noise during a class activity, and notes providing feedback or praise for individuals (Streimann et al., 2020). Additionally, the PAX version of GBG has its own distinct terminology, referring to rule-breaking behaviour as 'spleems' and desired behaviour as 'PAX behaviour' (Ialongo et al., 2019). Bradshaw et al. (2020) received a lower WoE C rating due to no explanation of how the intervention differed from the original. As previously stated, the efficacy of adapted or modified versions of the intervention are contingent upon adherence to the underlying principles of the GBG (Bowman-Perrot, 2016). As a result, the outcomes of Bradshaw et al.'s (2020) study should be interpreted with caution as it is unknown whether findings were attributed by the core principles of the original GBG.

In terms of intervention fidelity, six studies ensured that fidelity to the intervention was maintained. This involved providing mandatory training for teachers delivering the programme, as well as on-going mentoring from researchers to support the implementation of the GBG (Ashworth, 2020; Bradshaw, 2020; Breeman et al., 2016; Ialongo et al., 2019; Split et al., 2016; Streimann et al., 2020). Smith et al. (2018) reported that staff attended training but no mentoring was provided by researchers as on-going support. Hart et al. (2021) was the only study that did not report how fidelity to the intervention was maintained. As a result, lower fidelity scores were given to Smith et al. (2020) and Hart et al. (2021), which are reflected in WoE C ratings (see Appendix D, Table 2).

Measures

To measure social and behavioural outcomes in primary aged pupils, the Teacher Observation of Classroom Adaptation checklist (TOCA-C) and the revised scale (TOCA-R) were used in four studies (Ashworth et al., 2020; Bradshaw et al., 2020; Hart et al., 2021 & Ialongo et al., 2019). Authors reported the internal consistency across all subscales of the measure ($\alpha > .70$). Hart et al. (2021) and Ialongo et al. (2019) also used the Social Health Profile and independent observations of student's behaviour as additional measures for social and behavioural outcomes. This was accounted for in the WoE A ratings for using multiple measures (see Table 5).

Smith et al. (2018) and Streimann et al. (2020) used the Strengths and Difficulties Questionnaire (SDQ), with internal consistency reported for both teacher and parent rated scales ($\alpha > .60$). Both Breeman et al. (2016) and Split et al. (2016) used the Problem Behaviour at School Interview (PBSI). Breeman

et al. (2016) also used the Social Preference Procedure, whilst Split et al. (2016) used the Asocial Behaviour subscale of the Child Behaviour Scale and Peer Nomination Assessments as additional measures for social and behavioural outcomes. However, only the internal consistency was reported for the PBSI (α > .87) and the Child Behaviour Scale (α = .90). Finally, all eight studies did not report data on the validity of measures used, which is reflected in WoE A ratings (see Table 5).

Outcomes

Each study provided a variety of outcomes, with five studies reporting an effect size using Cohen's *d* (Ashworth et al., 2020; Bradshaw et al., 2020, Hart et al., 2021; Smith et al., 2018; Split et al., 2016) and three studies providing only inferential statistics, which was accounted for in WoE A ratings (see Table 5). In order to ensure consistency amongst studies, this review calculated the standardised mean difference (Cohen's *d*) to the latter studies so that effect sizes could be compared. Calculations were made using the Campbell Collaboration online calculator (Wilson, n.d.). The effect size for two studies were calculated by finding the pre- and post-test mean difference between intervention and comparison group, divided by the combined standard deviation (Breeman et al., 2016; Streimann et al., 2020). The effect size for Ialongo et al. (2019) was calculated by using the *F*-statistic of the intervention between pre- and post-test scores for both intervention and comparison group. See Table 6 for the outcomes and effect sizes for all eight studies.

Table 6.

Outcomes and Effect Sizes for Reviewed Studies

Study	Sample size	Relevant Measure(s)	Behaviour Outcome	Cohen's *d* [1]	Effect Size Descriptor	Significance value	95% CI	Overall Judgement Rating (WoED)
Ashworth et al. (2020)	N=3084	<u>Teacher rated:</u> Teacher Observation of Classroom Adaptation (TOCA-C)	Disruptive	*d* = 0.06	Negligible	*p* = .05	0.02 - 0.09	High
			Prosocial	*d* = -0.11	Negligible	*p* > .05	-0.32 - 0.10	
Bradshaw et al. (2020)	N=1526	<u>Teacher rated:</u> Teacher Observation of Classroom Adaptation-Revised (TOCA-R)	Hyperactivity	*d* = 0.00	Negligible	*p* > .05	-0.08 - 0.20	Medium
			Aggressive-disruptive	*d* = 0.00	Negligible	*p* > .05	-0.12 - 0.24	
			Social competence	*d* = 0.00	Negligible	*p* > .05	-0.14 - 0.22	
			Peer relations	*d* = 0.00	Negligible	*p* > .05	-0.08 - 0.24	
Breeman et al. (2016)	N=389	<u>Teacher reported:</u> Problem Behaviour at School Interview (PBSI)	Behavioural problems	*d* = 0.03	Negligible	*p* < .05*	-0.17 - 0.23	Low
		<u>Child reported:</u> Children's Social Preference Procedure	Social preference	*d* = 0.06	Negligible	*p* > .05	-0.14 - 0.27	

Study	Sample size	Relevant Measure(s)	Behaviour Outcome	Cohen's *d* [1]	Effect Size Descriptor	Significance value	95% CI	Overall Judgement Rating (WoED)
Hart et al. (2021)	N=650	<u>Teacher rated:</u> Teacher Observation of Classroom Adaption-Revised (TOCA-R)	Authority acceptance	*d* = 0.01	Negligible	*p* > .05	-0.22 - 0.15	Low
		<u>Researcher:</u> Independent behaviour observation	Total problem behaviour	*d* = 0.07	Negligible	*p* > .05	-0.26 - 0.12	
		<u>Teacher rated:</u> Social Health Profile	Social competence	*d* = 0.09	Negligible	*p* > .05	-0.01 - 0.27	
Ialongo et al. (2019)	N=5611	<u>Researcher:</u> Independent behaviour observation	Total problem behaviour	*d* = 0.08	Negligible	*p* < .05*	0.01 - 0.14	Medium
		<u>Teacher rated:</u> Teacher Observation of Classroom Adaptation-Revised (TOCA-R)	Authority acceptance	*d* = 0.01	Negligible	*p* > 0.05	-0.06 - 0.07	
		Social Health Profile (Social Competence subscale)	Social competence	*d* = 0.06	Negligible	*p* > .05	-0.00 - 0.13	

Study	Sample size	Relevant Measure(s)	Behaviour Outcome	Cohen's *d* [1]	Effect Size Descriptor	Significance value	95% CI	Overall Judgement Rating (WoED)
Smith et al. (2018)	N=811	<u>Child rated:</u> Strengths and Difficulties Questionnaire (SDQ)	Hyperactivity	*d* = -0.03	Negligible	*p* > .05	-0.13 - 0.07	Medium
			Prosocial	*d* = 0.08	Negligible	*p* < .05*	0.00 - 0.16	
			Conduct problems	*d* =-0.06	Negligible	*p* > .05	-0.16 - 0.04	
Split et al. (2016)	N=570	<u>Teacher rated:</u> Strengths and Difficulties Questionnaire (SDQ)	Hyperactive	*d* = 0.09	Negligible	*p* > .05	-0.24 -0.42	Medium
			Oppositional	*d* = 0.05	Negligible	*p* > .05	-0.22 - 0.33	
		<u>Child rated:</u> The Child Behaviour Scale (The Asocial Behaviour subscale)	Withdrawn from peers	*d* = 0.07	Negligible	*p* > .05	-0.01 - 0.15	
		<u>Child rated:</u> Peer nomination assessments	Hyperactive	*d* = -0.01	Negligible	p > .05	-0.04 - 0.02	
			Oppositional	*d* = -0.03	Negligible	p < .05*	-0.12 - 0.36	
			Withdrawn from peers	*d* = -0.01	Negligible	p > .05	-0.02 - 0.01	

Study	Sample size	Relevant Measure(s)	Behaviour Outcome	Cohen's *d* [1]	Effect Size Descriptor	Significance value	95% CI	Overall Judgement Rating (WoED)
Streimann et al. (2020)	N=708	<u>Teacher reported:</u> Strengths and Difficulties Questionnaire (SDQ)	Conduct problems	*d* = -0.07	Negligible	*p* < .05*	-0.22 - 0.07	Low
			Peer problems	*d* = 0.13	Negligible	*p* > .05	-0.01 - 0.28	
			Hyperactivity	*d* = 0.09	Negligible	*p* > .05	-0.05 - 0.25	
			Prosocial	*d* = -0.17	Negligible	*p* > .05	-0.30 - -0.01	
		<u>Parent reported:</u> Strengths and Difficulties Questionnaire (SDQ)	Conduct problems	*d* = -0.06	Negligible	*p* > .05	-0.23 - 0.12	
			Peer problems	*d* = 0.00	Negligible	*p* > .05	-0.18 - 0.18	
			Hyperactivity	*d* = 0.03	Negligible	*p* > .05	-0.21 - 0.15	
			Prosocial	*d* = 0.05	Negligible	*p* > .05	-0.13 - 0.22	

[1]According to Cohen (1988) an effect size considered negligible = < 0.2, small = 0.2, medium = 0.5 and large = 0.8
**indicates a significant effect less than or equal to 0.05.*

Four of the eight studies found that pupils in the intervention group demonstrated a significant reduction in problem behaviour at post-test, compared to the control group. However, the effect sizes were negligible (d < 0.2). Additionally, the type of problem behaviour varied considerably across studies. Split et al. (2016) found a reduction in pupil-reported oppositional behaviour but not in teacher-reported ratings. Additionally, non-significant effects were found for both teacher and pupil-reported measures for hyperactivity. This suggests that oppositional behaviours may be perceived differently by teachers and pupils. Streimann et al. (2020) found a significant reduction in teacher-reported conduct problems but not for hyperactivity. Two studies that used a general measure of behavioural problems (Breeman et al., 2016; Ialongo et al., 2019) found a significant reduction in behaviour at post-test compared to the control group. However, this was not the case for Hart et al. (2021) and Ashworth et al (2020) who used similar measures.

In regards to social behaviours, one study reported a significant improvement in prosocial behaviours (Smith et al., 2018), demonstrating a negligible effect size (d = 0.08). Non-significant effects were found for social competence, peer relations, withdrawn from peers and prosocial skill across the remaining studies.

Three studies did not report any significant effects for both behavioural and social outcomes (Ashworth, et al., 2020; Bradshaw et al., 2020; Hart et al., 2021). Bradshaw et al. (2020) and Hart et al. (2021) received lower WoE C ratings for fidelity (Hart et al., 2021) and origins of the intervention (Bradshaw et al., 2020), thus findings may account for these limitations (see Appendix D, Table 2 for WoE C ratings).

Despite some of the studies demonstrating significant differences between the intervention and control group, a negligible effect size indicates that the differences are too small to infer that improvements in social and behavioural outcomes are a result of the GBG (Thompson, 2007).

Conclusions and Recommendations

Discussion of Findings

The aim of this systematic literature review was to examine the effectiveness of the GBG on social and behavioural outcomes in primary aged pupils. Out of the eight studies appraised in this review, according to Gough's (2007) WoE Framework and Gersten's (2005) coding protocol, one study received a 'high' WoE D rating (Ashworth et al., 2020), three studies were given a 'low' rating, whilst the four remaining studies received a 'medium' rating. 'High' WoE D ratings, indicate studies of higher quality and should be given greater weight when interpreting this review. Overall, the findings indicate insufficient evidence to support that the GBG is an effective intervention for improving social and behavioural outcomes in primary aged pupils. Therefore, it is recommended that EPs should consider alternative interventions with a stronger evidence-base when supporting schools address concerns with social and behavioural difficulties.

The findings contrast with previous reviews and meta-analyses (Bowman-Perrot et al., 2016; Flower et al., 2014; Tingstrom et al., 2006) that reviewed the effects of the GBG from single-case experimental designs (SCDs), demonstrating significant and large effect sizes. A possible reason for this is that SCDs may inflate the effect size due to small sample sizes and the use of repeated

measures typically employed in such designs (Dunlap et al., 1996). Additionally, SCDs are not suitable for answering how effective an intervention is (Petticrew & Roberts, 2003). Therefore, this review provides more representative findings in regards to the effectiveness of the GBG.

Limitations and Recommendations

One study was conducted in the UK (Ashworth et al., 2020). Therefore, the findings are not generalisable to the UK population. Further research, using RCTs, should be conducted in the UK to determine whether similar findings are replicated; yielding greater generalisability of findings.

Three studies included participants who were identified as displaying higher levels of social and behavioural difficulties (Bradshaw et al., 2020; Breeman et al., 2016; Hart et al., 2021). Whilst these studies did not produce particularly convincing findings, three studies are not enough to reliably support previous assumptions that the effects of the GBG are more prominent for pupils experiencing higher problem behaviour at baseline (Breeman et al., 2016). Therefore, further research recruiting participants with higher problem behaviour is warranted.

Out of all the reviewed studies, Ashworth et al. (2020) was the only study to examine the difference between conditions by administering a survey to determine how the principles of the GBG differs from their usual behaviour management practices and procedures. Future studies should consider how usual provision conditions differ from the intervention to determine whether the findings can be attributed to the intervention.

Due to the flexibility the GBG provides in determining what behaviours to target contingent upon the specific needs of the class (Bowman-Perrott et al., 2016), little is known about what behaviours were being targeted within and between studies which may account for the overall findings. Future research should consider the specific behaviours being targeted during the intervention so that the relationship between the target behaviours and outcome measures can be closley examined.

Finally, the author of this review did not consider the sufficient power of the sample size as criteria for the WoE C. All eight studies included large sample sizes which may have been overpowered, potentially accounting for the effect size. Future reviews using Gough's (2007) WoE Framework should consider the power of the sample in their WoE C to further understand and evaluate the evidence.

References:

Ashworth, E., Humphrey, N., & Hennessey, A. (2020). Game Over? No Main or Subgroup Effects of the Good Behavior Game in a Randomized Trial in English Primary Schools. *Journal of Research on Educational Effectiveness*, *13*(2), 298–321. https://doi.org/http://dx.doi.org/10.1080/19345747.2019.1689592

Bandura, A. (1977). *Social learning theory*. Englewood Cliffs, NJ: Prentice Hall.

Barrish, H. H., Saunders, M., & Wolf, M. M. (1969). Good behavior game: Effects of individual contingencies for group consequences on disruptive behavior in a classroom 1. *Journal of Applied Behavioural Analysis*, *2*(2), 119-124.

Bowman-Perrott, L., Burke, M. D., Zaini, S., Zhang, N., & Vannest, K. (2016). Promoting Positive Behavior Using the Good Behavior Game: A Meta-Analysis of Single-Case Research. *Journal of Positive Behavior Interventions*, *18*(3), 180–190. https://doi.org/10.1177/1098300715592355

Bradshaw, C. P., Shukla, K. D., Pas, E. T., Berg, J. K., & Ialongo, N. S. (2020). Using Complier Average Causal Effect Estimation to Examine Student Outcomes of the PAX Good Behavior Game When Integrated with the PATHS Curriculum. *Administration and Policy in Mental Health and Mental Health Services Research*, *47*(6), 972–986. https://doi.org/10.1007/s10488-020-01034-1

Breeman, L. D., van Lier, P. A. C., Wubbels, T., Verhulst, F. C., van der Ende, J., Maras, A., Struiksma, A. J. C., Hopman, J. A. B., & Tick, N. T. (2016). Effects of the Good Behavior Game on the Behavioral, Emotional, and

Social Problems of Children with Psychiatric Disorders in Special Education Settings. *Journal of Positive Behavior Interventions*, *18*(3), 156–167. https://doi.org/http://dx.doi.org/10.1177/1098300715593466

Chan, G., Foxcroft, D., Smurthwaite, B., Coomes, L., & Allen, D. (2012). Improving child behaviour management: An evaluation of the Good Behaviour Game in UK primary schools. *Oxford: Oxford Brookes University.*

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

Coombes, L., Chan, G., Allen, D., & Foxcroft, D. R. (2016). Mixed-methods Evaluation of the Good Behaviour Game in English Primary Schools. *Journal of Community & Applied Social Psychology*, *26*(5), 369–387. https://doi.org/http://dx.doi.org/10.1002/casp.2268

Department for Education, & Department of Health (DfE & DoH). (2015). *Special educational needs and disability code of practice: 0 to 25 years.* GOV.UK. *https://www.gov.uk/government/publications/send-code-of-practice-0-to-25.*

Department of Education (DfE). (2021). *Permanent and fixed-period exclusions in England*. GOV.UK. https://explore-education-statistics.service.gov.uk/find-statistics/permanent-and-fixed-period-exclusions-in-england.

Department of Health and Social Care, & Department of Education (DHSC & DfE). (2017). *Transforming children and young people's mental health*

provision: a green paper. GOV.UK. https://www.gov.uk/government/consultations/transforming-children-and-young-peoples-mental-health-provision-a-green-paper.

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods*, *1*(2), 170.

Flower, A., McKenna, J. W., Bunuan, R. L., Muething, C. S., & Vega Jr, R. (2014). Effects of the Good Behavior Game on challenging behaviors in school settings. *Review of Educational Research*, *84*(4), 546-571.

Gersten, R., Fuchs, L. S., Compton, D., Coyne, M., Greenwood, C., & Innocenti, M. S. (2005). Quality indicators for group experimental and quasi-experimental research in special education. *Exceptional Children*, *71*(2), 149-164.

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.

Hart, S. R., Celene, D., Embry, D. D., Becker, K., Lawson, A., & Ialongo, N. (2021). The Effects of Two Elementary School-Based Universal Preventive Interventions on Special Education Students' Socioemotional Outcomes: RASE. *Remedial and Special Education*, *42*(1), 31–43. https://doi.org/http://dx.doi.org/10.1177/0741932520941603

Health & Care Professions Council (HCPC). (2016). *Standards of conduct, performance and ethics*. Health & Care Professions Council.

https://www.hcpc-uk.org/standards/standards-of-conduct-performance-and-ethics/.

Hukkelberg, S., Keles, S., Ogden, T., & Hammerstrøm, K. (2019). The relation between behavioral problems and social competence: A correlational Meta-analysis. *BioMed Central Psychiatry*, *19*(1), 1-14.

Humphrey, N., Hennessey, A., Ashworth, E., Frearson, K., Black, L., Petersen, K., ... & Pampaka, M. (2018). Good Behaviour Game. Evaluation Report and Executive Summary. *Education Endowment Foundation, London*.

Ialongo, N. S., Domitrovich, C., Embry, D., Greenberg, M., Lawson, A., Becker, K. D., & Bradshaw, C. (2019). A Randomized Controlled Trial of the Combination of Two School-Based Universal Preventive Interventions. *Developmental Psychology*, *55*(6), 1313–1325. https://doi.org/http://dx.doi.org/10.1037/dev0000715

Kellam, S. G., Brown, C. H., Poduska, J. M., Ialongo, N. S., Wang, W., Toyinbo, P., ... Wilcox, H. C. (2008). Effects of a universal classroom behavior management program in first and second grades on young adult behavioral, psychiatric, and social outcomes. Drug and Alcohol Dependence, 95 Suppl 1, S5–S28. http://doi.org/10.1016/j.drugalcdep.2008.01.004

Kellam, S. G., Mackenzie, A. C. L., Brown, C. H., Poduska, J. M., Wang, W., Petras, H., & Wilcox, H. C. (2011). The good behavior game and the future of prevention and treatment. *Addiction Science & Clinical Practice*, *6*(1), 73–84.

Leflot, G., van Lier, P. A., Onghena, P., & Colpin, H. (2010). The role of teacher behavior management in the development of disruptive behaviors: An intervention study with the good behavior game. *Journal of Abnormal Child Psychology*, *38*(6), 869-882.

Office for Standards in Education (Ofsted). (2014). *Below the radar: low-level disruption in the country's classrooms*. London: Ofsted.

Petticrew, M., & Roberts, H. (2003). Evidence, hierarchies, and typologies: horses for courses. *Journal of Epidemiology & Community Health*, *57*(7), 527-529.

Sewell, A. (2020). An Adaption of the Good Behaviour Game to Promote Social Skill Development at the Whole-Class Level. *Educational Psychology in Practice*, *36*(1), 93–109. https://doi.org/http://dx.doi.org/10.1080/02667363.2019.1695583

Skinner, B. F. (1945). The operational analysis of psychological terms. *Psychological Review*, *52*(5), 270.

Smith, E. P., Osgood, D. W., Oh, Y., & Caldwell, L. C. (2018). Promoting Afterschool Quality and Positive Youth Development: Cluster Randomized Trial of the Pax Good Behavior Game. *Prevention Science*, *19*(2), 159–173. https://doi.org/10.1007/s11121-017-0820-2

Spilt, J. L., Leflot, G., Onghena, P., & Colpin, H. (2016). Use of Praise and Reprimands as Critical Ingredients of Teacher Behavior Management: Effects on Children's Development in the Context of a Teacher-Mediated Classroom Intervention. *Prevention Science*, *17*(6), 732–742. https://doi.org/http://dx.doi.org/10.1007/s11121-016-0667-y

Streimann, K., Selart, A., Trummal, A., Karin, S., Anne, S., & Aire, T. (2020). Effectiveness of a Universal, Classroom-Based Preventive Intervention (PAX GBG) in Estonia: a Cluster-Randomized Controlled Trial. *Prevention Science*, *21*(2), 234–244. https://doi.org/http://dx.doi.org/10.1007/s11121-019-01050-0

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

Tingstrom, D. H., Sterling-Turner, H. E., & Wilczynski, S. M. (2006). The good behavior game: 1969-2002. *Behavior Modification*, *30*(2), 225-253.

Wilson, D. B. (n.d). *Practical Meta-Analysis Effect Size Calculator.* Campbell Collaboration. https://www.campbellcollaboration.org/research-resources/effect-size-calculator.html.

Appendices

Appendix A.

Table 1.

Studies excluded at full-text screening based on exclusion criteria

	Study reference	Exclusion Criteria number(s)
1.	Coombes, L., Chan, G., Allen, D., & Foxcroft, D. R. (2016). Mixed-methods evaluation of the good behaviour game in English primary schools. *Journal of community & applied social psychology*, *26*(5), 369-387.	4
2.	Donaldson, J. M., Wiskow, K. M., & Soto, P. L. (2015). Immediate and distal effects of the good behavior game. *Journal of Applied Behavior Analysis*, *48*(3), 685-689.	4
3.	Groves, E. A., & Austin, J. L. (2019). Does the Good Behavior Game Evoke Negative Peer Pressure? Analyses in Primary and Secondary Classrooms. *Journal of Applied Behavior Analysis*, *52*(1), 3–16.	4
4.	Lynne, S., Radley, K. C., Dart, E. H., Tingstrom, D. H., Barry, C. T., & Lum, J. D. K. (2017). Use of a technology-enhanced version of the good behavior game in an elementary school setting. *Psychology in the Schools*, *54*(9), 1049–1063.	4 & 5
5.	McHugh, D. M. B., Radley, K. C., Tingstrom, D. H., Dart, E. H., & Barry, C. T. (2019). The Effects of Tootling via ClassDojo on Pupil Behavior in Elementary Classrooms. *School Psychology Review*, *48*(1), 18–30.	4 & 5
6.	Ortiz, J., Bray, M. A., Bilias-Lolis, E., & Kehle, T. J. (2017). The Good Behavior Game for Latino English Language Learners in a Small-Group Setting. *International Journal of School & Educational Psychology*, *5*(1), 26–38.	4 & 5
7.	Sewell, A. (2020). An adaption of the Good Behaviour Game to promote social skill development at the whole-class level. *Educational Psychology in Practice*, *36*(1), 93-109.	4
8.	Sondey, J., Taurel, N., Khem, C., Negre, L., Birocchi, S., & Reynaud-Maurupt, C. (2019). The Good Behavior Game: when the classroom becomes the playground for life skills (Toulon area). *European Journal of Public Health*, *29*.	4
9.	Stratton, K. K., Gadke, D. L., & Morton, R. C. (2019). Using the Good Behavior Game with High School Special Education Pupils: Comparing Pupil- and Teacher-Selected Reinforcers. *Journal of Applied School Psychology*, *35*(2), 105–121.	4 & 5
10.	Torok, M., Rasmussen, V., Wong, Q., Werner-Seidler, A., Bridianne, O., Toumbourou, J., & Alison, C. (2019). Examining the impact of the Good Behaviour Game on emotional and behavioural problems in primary school children: A case for	4

	integrating well-being strategies into education. *Australian Journal of Education*, *63*(3), 292–306.	
11.	Wu, Y. Q., Chartier, M., Ly, G., Phanlouvong, A., Shelby, T., Weenusk, J., Murdock, N., Munro, G., & Sareen, J. (2019). Qualitative case study investigating PAX-good behaviour game in first nations communities: insight into school personnel's perspectives in implementing a whole school approach to promote youth mental health. *BMJ Open*, *9*(9).	4

Appendix B.

Example of WoE A Coding Protocol

Coding protocol: Gersten, R., Fuchs, L. S., Compton, D., Coyne, M., Greenwood, C., & Innocenti, M. S. (2005). Quality indicators for group experimental and quasi-experimental research in special education. *Exceptional Children*, *71*(2), 149-164.

Note. This protocol was adapted so that the questions are relevant to the research question. Question wording will be strikethrough (e.g. ~~example)~~ and appropriately re-worded if it is not relevant to the review. Explanations of adaptation will be identified through italics.

Reference of the study: Ashworth, E., Humphrey, N., & Hennessey, A. (2020). Game Over? No Main or Subgroup Effects of the Good Behavior Game in a Randomized Trial in English Primary Schools. *Journal of Research on Educational Effectiveness*, *13*(2), 298-321.

Essential Quality Indicators

A. Quality indicators for describing participants

Was sufficient information provided about the participants involved in the study? ~~to determine whether the participants demonstrated the difficulties presented?~~ *(the intervention is a universal programme so for the purpose of the review, this question will focus on whether the study has described the participant characteristics involved in the study)*

☒Yes

☐ No

☐ Unknown/Unable to Code

Were appropriate procedures used to increase the likelihood that relevant characteristics of participants in the sample were comparable across conditions?

☒ Yes

☐ No

☐ Unknown/Unable to Code

Was sufficient information given characterizing the interventionists or teachers provided? Did it indicate whether they were comparable across conditions?

☒ Yes; partially, Teachers allocated to intervention condition provided training and support to implement GBG.

☐ No

☐ Unknown/Unable to Code

B. Quality indicators for implementation of the intervention and description of comparison conditions

Was the intervention clearly described?

☒ Yes

☐ No

☐ Unknown/Unable to Code

Was the fidelity of implementation described and assessed?

☒Yes

☐ No

☐ Unknown/Unable to Code

Was the nature of services provided in comparison conditions described?

☒ Yes

☐ No

☐ Unknown/Unable to Code

C. Quality indicators for outcome measures

Were multiple measures used to provide an appropriate balance between measures closely aligned with the intervention and measures of generalised performance?

☐ Yes

☒ No; one measure assessed social and behavioural outcomes

☐ Unknown/Unable to Code

Were outcomes for capturing the intervention's effect measured at the appropriate times?

☒ Yes

☐ No

☐ Unknown/Unable to Code

D. Quality indicators for data analysis

Were the data analysis techniques appropriately linked to key research questions and hypotheses? Were they appropriately linked to the unit of analysis in the study?

☒ Yes

☐ No

☐ Unknown/Unable to Code

Did the research report include not only inferential statistics but also effect size calculations?

☒ Yes

☐ No

☐ Unknown/Unable to Code

Desirable Quality Indicators

Was data available on attrition rates among intervention samples?

☒ Yes

☐ No

☐ Unknown/Unable to Code

Was severe overall attrition (30% or more) avoided? Is attrition comparable across samples?

☒ Yes

☐ No

☐ Unknown/Unable to Code

Did the study provide not only internal consistency reliability but also test-retest reliability and interrater reliability (when appropriate) for outcome measures?

☐ Yes

☒ No; only internal consistency reliability reported (coefficient .87)

☐ Unknown/Unable to Code

Were data collectors and/or scorers blind to study conditions and equally (un)familiar to examinees across study conditions?

☐ Yes

☒ No

☐ Unknown/Unable to Code

Were outcomes for capturing the intervention's effect measured beyond an immediate post-test?

☐ Yes

☒ No

☐ Unknown/Unable to Code

Was evidence of the criterion-related validity and construct validity of the measures provided?

☐ Yes

☒ No

☐ Unknown/Unable to Code

Did the research team assess not only surface features of fidelity implementation (e.g. number of minutes allocated to the intervention or teacher/interventionist following procedures specified), but also examine quality of implementation?

☒ Yes

☐ No

☐ Unknown/Unable to Code

Was any documentation of the nature of instruction or series provided in comparison conditions?

☐ Yes

☒ No

☐ Unknown/Unable to Code

Did the research report include actual audio or videotape excerpts or examples of paperwork that capture the nature of the intervention?

☐ Yes – examples of materials used in all three groups

☒ No

☐ Unknown/Unable to Code

Were results presented in a clear, coherent fashion?

☒ Yes

☐ No

☐ Unknown/Unable to Code

Overall Rating of Evidence: ☒3 ☐ 2 ☐1 ☐ 0

WoE A Criteria – based on Gersten et al. (2005) coding protocol

		Low quality = 1 *(study meets less than 9 essential criteria)*	**Medium quality = 2** *(study meets at least 9 essential criteria AND at least 1 and less than 4 desirable criteria)*	**High quality = 3** *(Study meets at least 9 essential criteria AND 4 or more desirable criteria)*	**Over all rating (1-3)**
No. of essential quality indicators met	9/10			X	3
No. of desirable quality indicators met	4/8				

Appendix C.

Weight of Evidence B: Methodological Relevance

The WoE B determines how relevant the methodology used within each study is in answering the review question on the effectiveness of the Good Behaviour Game on social and behavioural outcomes for primary aged pupils. This was evaluated using Petticrew and Roberts (2003) typology of evidence criteria (see Table 1 and 2 for criteria and rationale). All criteria must be met to fulfil rating (see Table 3 for WoE B ratings).

Table 1.

WoE B Rating Criteria

WoE Rating	Criteria
High = 3	Design: • Randomised control trials with random assignment to GBG or control group • Appropriate control group
Medium = 2	Design: • Appropriate control group • Non-random assignment of participants
Low = 1	Design: • Single-case research • Qualitative research • No control group

Table 2.

Rationale for WoE B Criteria

Criteria	Rationale
Design	• Randomised control trials are identified as a high-quality research design to examine an interventions effectiveness (Petticrew & Roberts, 2003) • A comparison group allows for intervention effects to be compared against another intervention or no intervention

Table 3.

WoE B Rating for Reviewed Studies

Study	WoE B Rating
Ashworth et al. (2020)	3
Bradshaw et al. (2020)	3
Breeman et al. (2016)	3
Hart et al. (2020)	3
Ialongo et al. (2019)	3
Smith et al. (2018)	3
Split et al. (2016)	3
Streimann et al. (2020)	3

Appendix D.

Weight of Evidence C: Relevance to review question

The WoE C is a review-specific judgement that examines how relevant the study is in answering how effective the GBG is at improving social and behavioural outcomes for primary aged pupils (Gough, 2007). For this review, studies were appraised and given a WoE C rating outlined in table 1. Criteria were determined by the author of the review.

Table 1.

WoE C Criteria and Rationale

Criteria	WoE Rating	Descriptor	Rationale
Origins of intervention	3	The intervention is based on the original version of GBG	The research will be of high relevance to the review question if the intervention is based on the original principles of GBG (Barrish, Saunders & Wolf, 1969).
	2	The intervention is an adapted or enhanced version of GBG with an explanation of how it differs from the original	
	1	The intervention is an adapted or enhanced version of GBG without an explanation of how it differs from the original	
Setting	3	In schools in the United Kingdom	The research will be of relevance to Educational Psychologist's in the UK if the intervention has taken place in UK schools so that outcomes are generalisable to EP practice.
	2	In schools in OECD countries	
	1	In OECD countries in settings outside of school hours (e.g. youth clubs, afterschool centres)	
Fidelity	3	Teachers receive training and support from researchers	To ensure the intervention is maintaining

	2	Teachers receive training but no continued support from researchers	consistency across the sample the intervention should provide information on how fidelity to the intervention has been monitored
	1	Teachers do not receive training to implement the intervention	
Participants characteristics	3	Pupils displaying high level of disruptive behaviour	Existing research indicates that the intervention is most effective for pupils displaying disruptive behaviours at pre-test (Breeman, et al., 2015). Therefore, the relevance of the research should be delivered to pupils who are identified as displaying disruptive behaviour prior to the intervention
	2	Pupils not displaying high levels of disruptive behaviour	
	1	Not school aged pupils	

Table 2.

WoE C Rating Scores for Reviewed Studies

Studies	Origins of intervention	Setting	Fidelity	Participants	Overall WoE C
Ashworth (2020)	3	3	3	2	2.75 (High)
Bradshaw et al. (2020)	1	2	3	3	2.25 (High)
Breeman et al. (2016)	2	2	3	3	2.5 (High)
Hart et al. (2020)	2	2	1	3	2 (Medium)
Ialongo et al. (2019)	2	2	3	2	2.25 (High)
Smith et al. (2018)	2	1	2	2	1.75 (Medium)
Split et al. (2016)	2	2	3	2	2.25 (High)

Streimann et al. (2020)	2	2	3	2	2.25 (High)

Note: WoE C scores ≤ 1.5 = 'low', < 1.5 and ≥ 2.24 = 'medium', and ≥ 2.25 = 'high'