Case Study 1: An Evidence-Based Practice Review Report

Theme: School (setting) based interventions for children with special educational needs (SEN)

How effective is the Good Behaviour Game (GBG) at reducing displays of disruptive behaviour in secondary classroom settings?

Summary

The Good Behaviour Game (GBG; Barrish et al., 1969) is a group contingency procedure that has been widely assessed and heavily evaluated as a successful class-level intervention used to reduce disruptive behaviours in schools (Bowman-Perrott et al., 2016). Since its inception, there have been a variety of adaptations, and it has been widely used in primary school settings with little evidence to support its viability in secondary settings (Ford et al., 2020). This systematic review examined the effectiveness of the Good Behaviour Game at reducing displays of disruptive behaviour in secondary classroom settings. Five studies were included in this review based on a variety of inclusion criteria and were critically appraised through the use of Gough's (2007) Weight of Evidence Framework. The studies were then critically reviewed through two separate adapted coding protocols specific to their study design Horner et al. (2005) and Law et al. (1998). Effect sizes were mostly large, with one negligible effect found. Findings suggest that the Good Behaviour Game has a moderate effect at reducing disruptive behaviour within secondary classroom settings. Location and study design may contribute to the significance of this finding. Further research is suggested to explore the limitations identified by this review.

Introduction

The Good Behaviour Game

The Good Behaviour Game (GBG; Barrish et al., 1969) is a group contingency procedure that has been widely assessed and heavily evaluated. (Bowman-Perrott et al., 2016). It is an interdependent group contingency (Gresham & Gresham, 1982) and therefore, reinforcement to one member of a group relies on the behaviour of all members. The game takes place within the context of a classroom with the aim of encouraging pupils, both individually and in teams, to self-regulate their behaviour. Though many variations of the GBG have been evaluated, a few key elements have surfaced which aids the game's success. These are: having the children split into teams; the announcement of rules and expectations within the classroom; explaining the methods by which the team may win; positing points for violations (GBG response-cost) or acting in line with expectations (GBG reinforcement; Tanol et al., 2010); and providing reinforcement to those who earn points by meeting a predetermined criterion. The GBG was originally designed to be played for 10 minutes, for a frequency of three times per week which would steadily increase over the year (Kellam et al., 2011). It can be played daily and increased to the entire duration of a lesson. The aim of this is to reduce disruptive behaviours and increase pupils' motivation, interest, and academically engaging behaviours (Humphrey et al., 2018).

Psychological theory

The psychological underpinnings of the GBG can be understood through the behaviourist's principles of Operant conditioning, where behaviour was observed to be modified through the use of reinforcements or punishments

(Skinner, 1945). This theory highlights that desired behaviours can be encouraged through the use of rewards, as done in the game; and undesirable behaviours, for example disruptive behaviours, can be reduced by giving a punishment or sanction. Social Leaning Theory (Bandura, 1977) is also used within the GBG as the desired behaviours are modelled in the classroom by other pupils playing the game and this allows the opportunity for desired behaviours to be learned. Similarly Life course/Social field theory (Kellam et al., 2011) describes the classroom as a social field where pupils can learn to engage in behaviour that are appropriate according to the social task demand. Within the classroom that social demand would be to obey rules, socialise appropriately and focus, displays of such behaviour would indicate social adaptations which can be transferred to other contexts. The GBG would encourage this growth through both the teacher's instruction and peer modelling.

Rationale and Relevance

Disruptive behaviour negatively impacts both the students who are disruptive but also the rest of the students in the class as both learning opportunities and instruction time within the classroom are limited while teachers deal with the behaviour (Higgins et al., 2001). There is a correlation between rates of disruptive behaviour and methods of classroom management (Reinke et al., 2013). Disruptive behaviour in the classroom can diminish the quality of instruction and impact student grades and levels of participation (Ofsted, 2014). Therefore, effective intervention strategies are essential to ensure that both students and teachers thrive (Ford et al. 2020). Some of the areas in which group contingencies have proven to be successful are: decreasing

problematic behaviour; increasing pupils' demonstration of expected behaviour; completion of homework; increasing academic performance; controlling noise level and effecting categories of behaviour as a whole (Little et al., 2015).

This is particularly relevant in the United Kingdom as exclusion rates continue to be high despite being impacted by a global pandemic resulting in national lockdown. For 2019/ 2020, persistent disruptive behaviour remains as the most common cause of both permanent and fixed term exclusions, accounting for 34% each of the reasons exclusions are given (Department for Edcuation, 2021). Exclusion can impact the social, emotional, and mental wellbeing of pupils and can have long lasting effects. Teachers can experience stress and burnout as a result of consistently dealing with behavioural difficulties in the classroom and this can lead to teachers leaving the profession.

The above reasons illustrate a need for schools to be supported to implement either a class-wide or school-wide, evidence based behaviour management system which can promote appropriate social and academically engaging behaviours (Kellam et al., 2011). The Educational psychologist can help schools by helping them to think about how the GBG can help to reduce disruptive behaviours, can train school staff about how to deliver the intervention, support with implementation and can help to problem solve if any barriers arise as the intervention is being used. This can in turn help to alleviate some of the social, behaviours and academic challenges that disruptive behaviours may cause. Previous reviews have found promising

results for the positive impact that the GBG has on disruptive behaviour, but there is a limited evidence base for its use in secondary setting (Joslyn et al., 2019; Smith et al., 2021). The provides an adequate rationale for further exploration about the impact the GBG can have on the disruptive behaviour displayed in secondary school settings and this review provides this insight.

Review Question

How effective is the Good Behaviour Game (GBG) at reducing displays of disruptive behaviour in secondary classrooms settings?

Critical Review of the Evidence Base

Literature Search

Between January and February 2022, a systematic literature search was conducted to find research regarding the use of the Good Behaviour Game as an intervention within secondary settings. The search terms in Table 1 were used on the Educational Resources Information Centre (ERIC), PsycINFO, and OVID (Medline) databases.

Table 1.

Search Terms

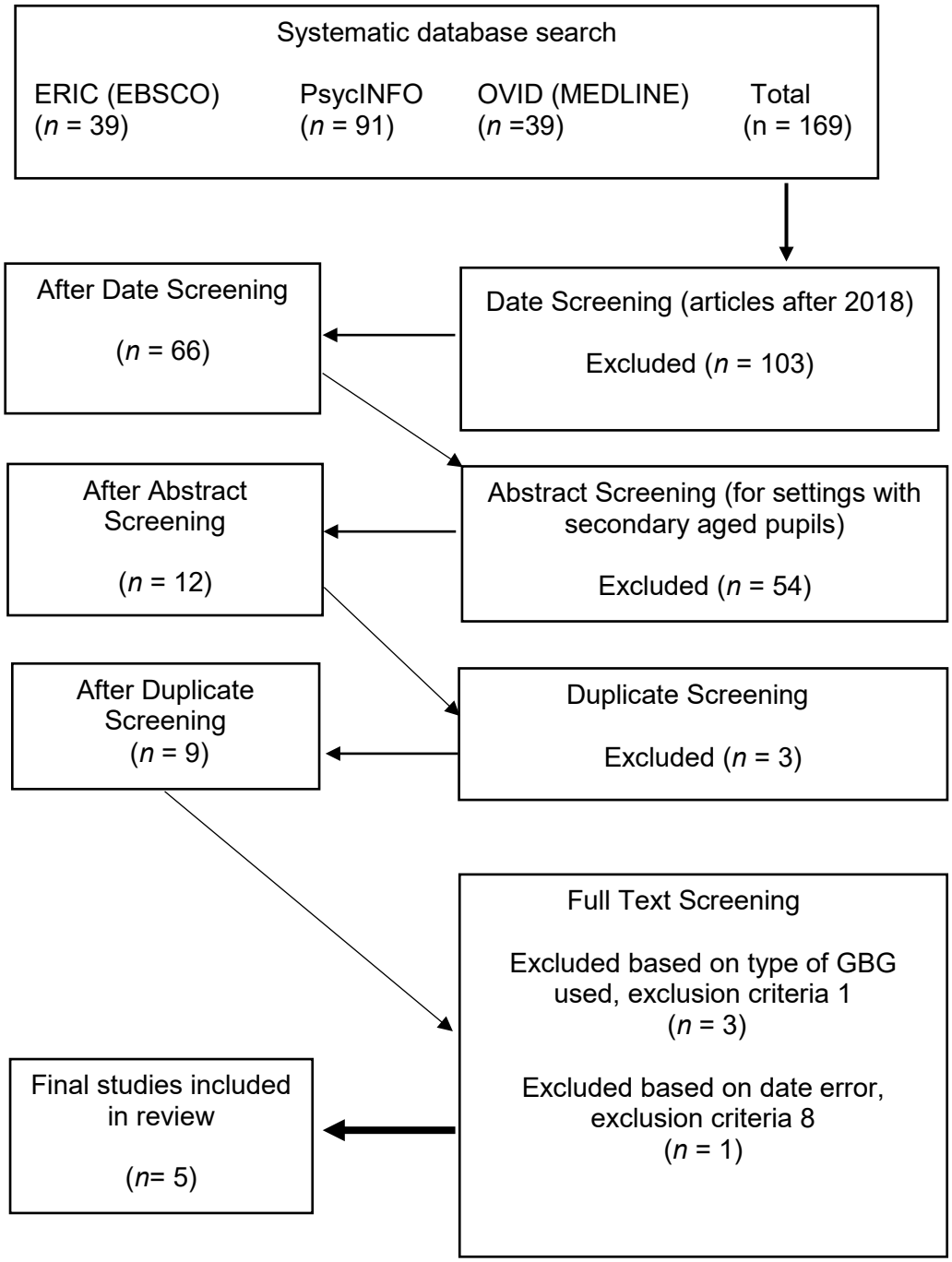
Intervention		Need		Context
good behavi* game	AND	challenging behavi* or disruptive behavi*	AND	classroom or school or learning environment

Screening Process

Across the three databases, a total of 169 studies were produced for screening. To ensure that the studies were relevant and current, they were screened through the use of predetermined inclusion and exclusion criteria which can be seen in Table 2. The last review including secondary aged pupils was conducted in 2018 and therefore the 1st level of screening began with limiting the studies to 2018 and beyond, this removed 103 studies. The participant focus of this review are secondary aged pupils, and this was screened for at the abstract level, which removed 54 studies. After removing duplicates, 9 studies remained for full text screening. A further 4 studies were removed based on the exclusion criteria pertaining to the intervention type and date, leaving 5 studies to be utilised for this review (Table 3). The excluded studies from this stage of the screening process can be found in Appendix A. Figure 1 is a graphic representation of the screening process as described above.

Figure 1.

Flow Chart Illustrating the Process of the Systematic Literature Search



Inclusion and Exclusion

Table 2.

Inclusion and Exclusion for Study Screening

Feature of Study	Inclusion Criteria	Exclusion Criteria	Rationale
1. Intervention	Good Behaviour Game (GBG) used with or without adapted elements, or technology facilitated versions of the GBG	Studies where the Good Behaviour Game is not used or the Caught Being Good Game (CBGG) is used	This review intends to explore the effectiveness of the GBG intervention.
2. Participants	Studies including pupils within secondary school or mixed primary/secondary school settings or making comparisons between primary and secondary pupils	Studies with pupils not in or making a comparison with pupils in secondary settings	This review aims to explore the effectiveness of the GBG intervention on pupils within secondary school settings.
3. Publication type	Studies which have been peer reviewed	Non-peer reviewed studies	To ensure studies have been reviewed meticulously.
4. Research design	Quantitative studies preferably experimental designs with comparison features	Qualitative studies, or experimental studies without comparison features	The comparative elements of an experimental design will support the review's aim to explore if any differences in outcome can be attributed to the effects of the GBG.
5. Outcomes	Measures behavioural outcomes pre- and post-intervention	Studies do not report measures for behavioural outcomes pre- and post- intervention	This review seeks to explore the effects of the GBG on the behavioural outcomes of pupils.
6. Location	Studies conducted in countries within the Organisation for Economic Co-Operation and Development (OECD)	Studies conducted in countries without OECD membership	The findings from studies conducted within OCED countries would be considered to be more generalisable to the UK population due to having similar education, demographic, and economic systems.
7. Language	Studies published in English	Studies published in languages besides English	Translation software not available to facilitate review of other languages.

8. Date of publication	Studies published in beyond 2018	Studies published prior to 2018	This review aims to examine current studies published about the GBG, and 2018 was the last date a review was conducted including secondary aged pupils.

Studies Included in Review

Table 3.

Reference List of Five studies

Studies Included in the Review
Ford, W. B., Radley, K. C., Tingstrom, D. H., & Dufrene, B. A. (2020). Efficacy of a No-Team Version of the Good Behaviour Game in High School Classrooms. *Journal of Positive Behaviour Interventions*, 22(3), 181–190. https://doi.org/10.1177/1098300719890059
Groves, E. A., & Austin, J. L. (2019). Does the Good Behaviour Game evoke negative peer pressure? Analyses in primary and secondary classrooms. *Journal of Applied Behaviour Analysis*, 52(1), 3–16. https://doi.org/10.1002/JABA.513
Stratton, K. K., Gadke, D. L., & Morton, R. C. (2019). Using the Good Behaviour Game with High School Special Education Students: Comparing Student- and Teacher-Selected Reinforcers. *Journal of Applied School Psychology*, 35(2), 105–121. https://doi.org/10.1080/15377903.2018.1509920
Troncoso, P., & Humphrey, N. (2021). Playing the long game: A multivariate multilevel non-linear growth curve model of long-term effects in a randomized trial of the Good Behaviour Game. *Journal of School Psychology*, 88, 68.
Vargo, K., & Brown, C. (2020). An evaluation of and preference for variations of the Good Behaviour Game with students with autism. *Behavioural Interventions*, 35(4), 560–570. https://doi.org/10.1002/BIN.1740

Mapping the Field

After conducting a systematic literature search, five studies were identified that described the effects of the Good Behaviour Game on disruptive behaviours in secondary aged pupils. The key features for each of these studies are detailed in Table 4.

Table 4.

Mapping the Field

Authors	Location	Sample & Participant Characteristics	Study Type & Control Group	Measures	Outcomes
Ford et al. (2020)	USA	Total N = **74** pupils from 3 schools **Class 1** (yr. 11)- N=27, **Class 2** (yr. 9) - N=21, **Class 3** (yr. 10) - N=26, with history of significant DB Setting: Mainstream Secondary School Student Ethnicity: **Class 1** – 21 African American & 6 Caucasian **Class 2** – 14 African American & 7 Caucasian **Class 3** – 25 African American & 1 Hispanic No gender data presented	Single case ABAB withdrawal design (Baseline, treatment, treatment withdrawn, reimplementation of treatment)	> Interval Observation > modified Behaviour Intervention Rating Scale (BIRS)	**Classroom 1** - Baseline: DB & AEB - (M = 41% and M = 33%; respectively); GBG: (M = 16%; and M = 52%, respectively); GBG withdrawn: (M = 28%; and M = 28%, respectively); and GBG reintroduced: (M = 10%; and M = 53%, respectively). **Classroom 2** - Baseline: DB & AEB - (M = 43% and M = 37%; respectively); GBG: (M = 14%; and M = 37%, respectively); GBG withdrawn: (M = 48%; and M = 18%, respectively); and GBG reintroduced: (M = 17%; and M = 39%, respectively). **Classroom 3** - Baseline: DB & AEB - (M = 41% and M = 20%; respectively); GBG: (M = 14%;

Key – Disruptive Behaviour (DB); Academically Engaged Behaviour (AEB); Good Behaviour Game (GBG); Mean (M)

	and M = 49%, respectively); GBG withdrawn: (M = 46%; and M = 18%, respectively); and GBG reintroduced: (M = 21%; and M = 41%, respectively).

Key – Disruptive Behaviour (DB); Academically Engaged Behaviour (AEB); Good Behaviour Game (GBG); Mean (M)

Stratton et al. (2019)	USA	Total N = **5** pupils (4 boys and 1 girl). Each student with unique learning needs and an Individualised Education Plan. **Team A** - N=2 **Team B** - N=3 Setting: Secondary School Resource Classroom Student Ethnicity: 5 African Americans No age or year data presented	Single case multielement withdrawal (A/[B+C]/A/[B+C]) design	> Visual analysis procedures > Usage Rating Profile-Intervention (URP-I)	**Team A** - Baseline: M = 6.67; GBG: M = 2.2. Reversal: M =1.6; GBG reintroduced: M = 0. **Team B** - Baseline: M = 8; GBG: M = 1.6. Reversal: M = 2.2; GBG reintroduced: M = 0.33. Data suggest that the GBG was an effective intervention for decreasing DB in the high school resource classroom across both teams. There was no difference in DB across different reward topographies.

Key – Disruptive Behaviour (DB); Good Behaviour Game (GBG); Mean (M)

Groves & Austin (2019)	Wales	Total N = **13** pupils from 2 schools – Setting: Secondary Pupil Referral Unit **Classroom 1** - N=5 (2 females, 3 males), aged 15-16, excluded from mainstream education due to excessive behavioural difficulties with 2 pupils diagnosed with a specific learning difficulty (they operated as 1 team). Setting: Special primary/secondary School **Classroom 2** - N=8 (2 females, 6 males), aged 9-10, all diagnosed with either global developmental delays, intellectual disabilities, or autism (they operated as 3 teams). No ethnicity data presented	Single case ABAB withdrawal design (Baseline, treatment, treatment withdrawn, reimplementation of treatment)	> Interval and Time sampling Observation > Teacher & Student Likert type questionnaire - Teacher's Social Validity Questionnaire & Students' Social Validity Questionnaires	**Classroom 1** - Baseline: off task behaviour - high (M = 66%); GBG: (M = 40%); GBG withdrawn: (M = 63%); and GBG reintroduced: (M = 11%). **Classroom 2** - Baseline: moderate levels of verbal and physical disruption - (M = 36% and M = 27%; respectively); GBG: (M = 9%; and M = 4%, respectively); GBG withdrawn: (M = 40%; and M = 28%, respectively); and GBG reintroduced: (M = 7%; and M = 5%, respectively). Reduced disruption in each classroom and improved peer relationships were noted by teachers and students as major changes resulting from the GBG intervention. both teachers and most students felt that the game was fair as measured by the social validity assessment.

Key – Disruptive Behaviour (DB); Good Behaviour Game (GBG); Mean (M)

Vargo & Brown (2020)	USA	Total N = 6 pupils (Males) aged 14 -16 with autism diagnosis Setting: Special Education Secondary Classroom Participants selected because they engaged in disruptive behaviour that prevent them from full participation in a general education classroom. No gender or ethnicity data presented	Single case multielement reversal design	> Interval Observation > Group-oriented concurrent-chains arrangement to assess student preference of variation	**Baseline phase** - M = 55%; **intervention phase**: Traditional GBG - M = 7%, ClassDoJo GBG - M= 4%, ClassBadges GBG - M = 7%. **Reversal Baseline** - M = 55%; **intervention phase**: Traditional GBG - M = 6%, ClassDoJo GBG - M= 2%, ClassBadges GBG - M = 8%. *During the Preference voting intervention phase* - ClassDojo GBG - M=2%. *During preference lottery intervention phase* - ClassDoJo and Traditional GBG used M = 3. Results showed that all three GBG variations were similarly effective in decreasing disruptive behaviours.

Key – Disruptive Behaviour (DB); Good Behaviour Game (GBG); Mean (M)

Troncoso & Humphrey (2021)	England	Total N = 3084 pupils (77 schools) middle childhood (ages 6-7 years) to early adolescence (ages 10-11 years) GBG at T1, N 1498 Control at T1, N 1469 GBG at T5, N 1051 Control at T5, N 1164 No gender or ethnicity data provided, however they mentioned, "intervention and control schools did not differ significantly with respect to sex, free school meals eligibility (FSM), English as an additional language (EAL), or special educational needs and disabilities (SEND; Humphrey et al., 2018)".	Multivariate multilevel non-linear growth curve model in a cluster randomized controlled trial (RCT)	>Teacher Observation of Child Adaptation Checklist > Structured Observation Schedule > Strength and Difficulties Questionnaire (SDQ)	No intervention effects were unequivocally found in relation to disruptive behaviour. Baseline pre-randomisation (T1), T2, T3, T4, and T5 (GBG and no intervention - control group split). ***Disruptive behaviour results:*** Control - **C** GBG - **G** *Implementation Phase* T1 – **C:** M = 1.612 SD = 0.812 T1- **G:** M = 1.709 SD = 0.810 T2 – **C:** M = 1.644 SD = 0.745 T2 - **G:** M = 1.761 SD = 0.798 T3 – **C:** M = 1.647 SD = 0.837 T3 - **G:** M = 1.740 SD = 0.856 *Follow-up Phase* T4 – **C:** M = 1.706 SD = 0.789 T4 - **G:** M = 1.747 SD = 0.854 T5 – **C:** M = 1.740 SD = 0.863 T5 - **G:** M = 1.732 SD = 0.840

Key – Trial Number (T); Control Group (C); Good Behaviour Game Intervention Group (G); Mean (M); Standard Deviation (SD)

Critical Appraisal of Included Studies

Weight of Evidence (WoE)

Gough's (2007) Weight of Evidence (WoE) Framework was used to critically assess to what extent the five reviewed studies answered the review question, based on the three dimensions measured. After careful review of each study, judgements were made about the methodological quality of the study (WoE A); the study's methodological relevance towards the question being reviewed (WoE B); and the topic relevance (WoE C).

For WoE A (Appendix B) each study was examined using a coding protocol suitable for their specific research design, to determine how closely the studies followed scientific methodological principles. The Troncoso and Humphrey (2021) study, was appraised through the use of criteria (Appendix B, Table 3b) from the Law et al. (1998) protocol. This was used because it is effective at evaluating quantitative studies (see Appendix F for the Law et al., (1998) version of the coding protocol). The protocol designed to appraise single-subject or small-N designs produced by Horner et al. (2005), was utilised to evaluate the methodological quality of four studies (Ford et al., 2020; Stratton et al., 2019; Groves & Austin, 2019; Vargo & Brown, 2020). An example of this can be found in Appendix E, and the criteria used is in Table 3a (Appendix B). More details about the criteria and rationale behind the ratings for WoE B and C can be found in Appendices C and D respectively.

An overall weight of evidence judgment (WoE D) for each study was calculated by averaging the evaluations made within each category (WoE A-C), these ratings can be found in Table 5.

Table 5.

Weight of Evidence - Overall Ratings

Authors	WoE A – Methodological Quality	WoE B – Methodological Relevance	WoE C – Topic Relevance	WoE D – Overall Score
Ford et al. (2020)	2.6 (High)	2 (Medium)	2.3 (Medium)	2.30 (Medium)
Stratton et al. (2019)	2.4 (Medium)	2 (Medium)	2.3 (Medium)	2.25 (Medium)
Groves & Austin (2019)	2.4 (Medium)	2 (Medium)	2.7 (High)	2.37 (Medium)
Vargo & Brown (2020)	2.9 (High)	2 (Medium)	2.3 (Medium)	2.40 (High)
Troncoso & Humphrey (2021)	2.6 (High)	3 (High)	2.7 (High)	2.75 (High)

Rating Key: High > 2.4, Medium = 1.5 - 2.4, Low = <1.4

Participants

Identifiable demographic information such as gender, ethnicity, and age were not reported consistently across the studies, which prevented the reviewer from identifying any patterns within the sample; This was reflected in the rating of WoE A (Appendix B).

Across all of the studies, the sample size of participants ranged from 5 to 3084. In total, 3182 pupils aged 6 - 16 years old were included in the reviewed studies. However, as this review gave attention to the effects of the GBG on adolescents, the more accurate description of the age range focused on is 10 - 16 years of age. This can be explained as Troncoso and

Humphrey's (2021) follow-up phase which measured the effects of the GBG on 10 to 11-year-olds in secondary schools in accordance with the inclusion criteria (see Table 2, criteria 2).

There were only three studies (Groves & Austin, 2019; Stratton et al., 2019; Vargo & Brown, 2020) that detailed gender distribution within their sample, and no pattern emerged in how the genders were split. Two studies (Ford et al., 2020; Stratton et al., 2019) reported ethnicity information, which showed a high concentration of African American participants within those studies; this cannot be interpreted as a trend for the reviewed sample as a whole.

In four of the five studies (Ford et al., 2020; Groves & Austin, 2019; Troncoso & Humphrey, 2021; Vargo & Brown, 2020), disruptive behaviours were described; three of the four (Ford et al., 2020; Groves & Austin, 2019; Vargo & Brown, 2020) indicated high levels, which were captured within their WoE C participant ratings as this review is intended to assess whether the usage of the GBG is associated with a decrease in disruptive behaviours (see Appendix D). As part of their studies (Groves & Austin, 2019; Stratton et al., 2019; Vargo & Brown, 2020), three authors recruited participants with special educational needs or diagnosed conditions, for example autism, global developmental delays, specific learning or intellectual disabilities, and autism (see Table 4 for SEN breakdown); each study identified a correlation between these conditions and the manifestation of disruptive behaviour.

Although Troncoso and Humphrey (2021) did not provide a detailed breakdown of demographic characteristics, they expressed that their sample reflects that of the typical UK school population; and they used cluster

randomisation to assign participants to experimental or control groups at the school level, rather than at the class level, thus minimising contamination risks and enhancing the validity of the study. This was considered within the WoE B rating (Appendix C).

Setting

In this review, two studies came from the United Kingdom (Groves & Austin, 2019; Troncoso & Humphrey, 2021) which resulted in a higher WoE C rating, because their results are more generalisable to UK schools and are therefore more relevant in applicability to the educational psychology field.

The studies within this review had a wide range of settings: mainstream secondary schools (Ford et al., 2020), a secondary pupil referral unit and a special primary/secondary school (Groves & Austin, 2019), a special education secondary classroom (Vargo & Brown, 2020), and a comparison between the transition from mainstream primary to mainstream secondary school (Troncoso & Humphrey, 2021). Stratton et al. (2019) was the only study that did not describe their setting to a degree that was clear enough to be replicated and therefore received a low rating in WoE A (see Appendix B).

Research design

All five studies are considered to be experimental designs though they vary in type. Troncoso and Humphrey (2021) were awarded the highest weighting for methodological relevance (WoE B) due to their use of random allocation of participants across the control and intervention conditions within their study. Each of the remaining four studies employed a single case design consisting of multiple baselines and treatment phases. These were used to

act as within-participant controls. Further experimental control was attained by the withdrawal or reversal of the intervention (Horner et al., 2005), which each study did. Using the participants to act as their own control demonstrated a clearer causal relationship between the intervention and behaviour change and gained each of these studies a WoE B score of medium.

Intervention

Each of the reviewed studies implemented the good behaviour game (GBG) in a variety of ways. Only two studies (Stratton et al., 2019; Troncoso & Humphrey, 2021) utilised the original version of the GBG without any modifications or added features. It is for this reason that both studies have received the highest intervention rating for WoE C. The other variations of the GBG included: a no teams version which promoted class wide participation (Ford et al., 2020); a comparison between a teamed and no team version (Groves & Austin, 2019); and a comparison between two technologically enhanced versions versus the original GBG. Due to the differences between these studies' interventions and the original version, they were given a lower WoE C rating, and this is because adapted or modified versions of the interventions will only be effective if they adhere to the GBG principles (Bowman-Perrot, 2016).

All of the studies reported on treatment and procedural integrity ranging from 70 – 100% across all phases and detailed how staff training was handled at the beginning and throughout the duration of the studies. This impacted the fidelity of the interventions used and added to its effectiveness.

Measures

All studies within this review utilised direct and structured observation measures to record the disruptive behaviours of participants.

Ford et al. (2020) utilised interval sampling observation tools for a duration of 20 minutes at 10 second intervals, this was done over a period of 5 consecutive days. The interobserver agreement (IOA) across all 3 classes ranged from 90.5% – 94%, this demonstrates good reliability as the acceptable minimum is 85% (Ford et al., 2020). Similarly, Vargo and Brown (2020) utilised interval sampling over a 5-day period but their observation was for 40 minutes at a 30 second interval. Their mean IOA was 95%, ranging from 89% - 100% across participants and phases. Groves and Austin (2019) used both interval and time sampling to measure disruptive behaviours once per day, 3 or 4 times per week for the duration of 15 – 30 minutes in 10 – 15 second intervals. IOA ranged from 89% - 96% across phases and teams. The final two studies were not as detailed regarding how observation was used but Stratton et al. (2019) mentioned the use of 49 minutes of observation over a 26-day period,18 of which were intervention days; and recorded an IOA of 86% and 97.6% for both teams. Though this reflects good reliability, the lack of clarity was reflected in the rating for WoE A. Troncoso and Humphrey (2021) utilised structured observations however that was for the purpose of measuring the procedural integrity of the intervention being carried out by the teachers. Instead, they used informant observational checklists and questionnaire to garner data about disruptive

behaviour patterns. These were the Teacher Observation of Child Adaptation Checklist (TOCA-C) and the Strength and Difficulties Questionnaire (SDQ), their use of multiple measures was reflected in WoE A rating as high.

Outcomes

Inferential statistics such as means and standard deviations were reported in all studies. Two studies reported non-overlap of all pairs (NAP; Ford et al., 2020; Stratton et al., 2019) following data analysis. NAP is a non-parametric analysis of paired data which is a comparison of baseline/withdrawal data points with subsequent intervention data points (College Station TX: Texas A&M University, n.d.). The reviewer calculated the NAP for the other two single case design studies which were missing (Groves & Austin, 2019; Vargo & Brown, 2020) in order to ensure consistency; this lack of data was reflected in their WoE A rating. This was achieved by uploading an image of the plotted graph into a website (Ankit Rohatgi, 2017) that gave me the original data points for each phase; these data points were then uploaded into another website (College Station TX: Texas A&M University, n.d.) that calculated the NAP for each phase, converting them into a final NAP value for the teams and classrooms represented.

Troncoso and Humphrey (2021) reported inferential statistics, and Cohen's *d* is typically used to measure effect sizes for randomised control trials (RCTs); Cohen, 2013). This reviewer converted the data to Cohen's *d* using the Campbell Collaboration online calculator (Wilson, n.d.). Only the effect size of the final phase was reported because it represented the most current impact of the GBG, that was measured whilst the pupils were in secondary school. It

is important to note that NAP is not comparable with Cohen's *d*. All five effect sizes are reported alongside their overall WoE D rating in Table 6.

Table 6.

Outcomes and Effect Sizes for the Reviewed Studies

Authors	Sample Size	Relevant Measures	Outcome	Effect Size	Effect size provided?	Description of Effect	Weight of Evidence D
Ford et al. (2020)	N = 74	> Interval Observation > modified Behaviour Intervention Rating Scale (BIRS)	The no-team version of the GBG was found to be effective in reducing levels of DB in regular education high school classrooms.	**Classroom 1** - Strong effect in reducing levels of DB (NAP = 1.00 or 100%) **Classroom 2** - DB are characterised as a strong effect (NAP = 1.00 or 100%). **Classroom 3** - The improvements in DB (NAP = 1.00 or 100%) is considered strong.	Yes, it was provided by the study	Classroom 1 - Strong Classroom 2 - Strong Classroom 3 - Strong	2.30 Medium
Stratton et al. (2019)	N = 5	> Visual analysis procedures > Usage Rating Profile-Intervention (URP-I)	Data suggest that the GBG was an effective intervention for decreasing DB in the high school resource classroom across both teams. There was no	**Team A** - Baseline & GBG NAP = 87.8% moderate effect. Reversal & GBG reintroduced: NAP = 80% moderate effect. **Team B** - Baseline & GBG NAP = 98.9% strong effect. Reversal & GBG reintroduced NAP = 93.3% strong effect.	Yes, it was provided by the study	Team A - Moderate Team B - Strong	2.25 Medium

			difference in DB across different reward topographies.				
Groves & Austin (2019)	N = 13	> Interval and Time sampling Observation > Teacher & Student Likert type questionnaire - Teacher's Social Validity Questionnaire & Students' Social Validity Questionnaires	Reduced disruption in each classroom and improved peer relationships were noted by teachers and students as major changes resulting from the GBG intervention. both teachers and most students felt that the game was fair as measured by the social validity assessment.	**Classroom 1** - both phases (Baseline 1 vs GBG 1, Baseline 2 vs GBG 2) reflect strong effects of NAP = 0.96 (96%) and 1.00 (100%) respectively for off-task behaviour. Combined NAP = 0.98 (98%) **Classroom 2** - both phases (Baseline 1 vs GBG 1, Baseline 2 vs GBG 2) reflect strong effects of NAP = 1.00 (100%) respectively for both verbal and physical disruptions. Combined NAP = 1.00 (100%)	No, it was calculated by the reviewer	Classroom 1 - Strong Classroom 2 - Strong	2.37 Medium

Vargo & Brown (2020)	N = 6	> Interval Observation > Group-oriented concurrent-chains arrangement to assess student preference of variation	Results showed that all three GBG variations were similarly effective at decreasing disruptive behaviours in secondary school setting.	All phase comparisons Baseline 1 vs GBG 1, Baseline 2 vs GBG 2, Baseline 2 vs Preference voting, Baseline 2 vs Preference Lottery reflect a strong effect NAP = 1.00 or 100%	No, it was calculated by the reviewer	Strong	2.40 High
Troncoso & Humphrey (2021)	N = 3084	>Teacher Observation of Child Adaptation Checklist structured > Observation Schedule > Strength and Difficulties Questionnaire (SDQ)	No intervention effects were unequivocally found in relation to disruptive behaviour.	only the effect of the final phase is being reported because this represents the most current impact that was measured whilst pupils were in secondary school *d* = 0.0094 (no effect)	Yes, it was provided by the study but the Cohen's *d* for the relevant phase was calculated by the reviewer	Negligible	2.75 High

Key: NAP can be reported as 1 or 100%; Strong effect = 0.93 – 1.0, Moderate effect = 0.66 – 0.92, Weak effect = 0 – 0.65 (Parker & Vannest, 2009). Cohen's *d* - < 0.2 = Negligible, Small effect = 0.2, Medium effect = 0.5, Large effect = 0.8 (Cohen, 2013).

Conclusion and Recommendations

Discussion of Findings

With strong effect sizes observed between the two classrooms, Ford et al. (2020) found that the no-team version of the GBG showed positive results in reducing disruptive behaviours in a mainstream secondary school setting. This trend was also found to be true across the variety of settings examined by this review, regardless of which version of the GBG was used. Strong effect sizes indicated a reduction of disruptive behaviour in the secondary pupil referral unit and special primary/secondary school even with the use of a teamed and no teamed version of the game (Groves & Austin, 2019); and a special education secondary classroom with original GBG used as well as technologically enhanced versions (Vargo & Brown, 2020). With a mixture of moderate and strong effects Stratton et al. (2019) found that disruptive behaviours reduced through the use of the original version of the GBG within a secondary resource base classroom. Due to the lack of statistically significant effects found by Troncoso and Humphrey (2021), there is no evidence to suggest that the GBG impacted disruptive behaviour for secondary pupils, after the game was played in primary school.

Conclusions and Limitations

This systematic review set out to examine the effectiveness of the Good Behaviour Game at reducing displays of disruptive behaviour in secondary classroom settings. The five studies that were included in this review were critically appraised through the use of Gough's (2007) Weight of Evidence Framework. The studies were then further appraised with two separate adapted coding protocols specific to their study design Horner et al. (2005)

and Law et al. (1998). Effect sizes were mostly large with one negligible effect found. These findings suggest that the Good Behaviour Game has a moderate effect at reducing disruptive behaviour within secondary classroom settings. However, given that majority of the studies reviewed utilised a single case design, with a small number of participants, this could be why strong effect sizes are reflected in the results. Also, these types of study designs are not seen as the strongest in answering the question of effectiveness and this is seen as a limitation of this review. As a result, within Educational Psychology practice, the Good Behaviour Game would not be my first recommended strategy to minimise disruptive behaviour in secondary schools. I would recommend any other behaviour management intervention with a stronger evidence base. Having said that, I would potentially recommend the use of the GBG in small classrooms, as this review does show promising evidence of its usefulness in that type of setting. I also believe that this review adds to the evidence base for the GBG's use with pupils with special educational needs.

The majority of the studies within this review were based within the United States of America and this made the findings lack generalisability to the United Kingdom. Further exploration is recommended to aptly examine the use of the GBG within the UK with a focus on secondary school populations.

References and Appendices

References

Ankit Rohatgi. (2017). *WebPlotDigitizer - Copyright 2010-2017 Ankit Rohatgi*. https://apps.automeris.io/wpd/

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

Barrish, H. H., Saunders, M., & Wolf, M. W. (1969). Good Behaviour Game: Effects of individual contingencies for group consequences on disruptive behaviour in a classroom. *Journal of Applied Behaviour Analysis*, *2*, 119–124

Bowman-Perrott, L., Burke, M. D., Zaini, S., Zhang, N., & Vannest, K. (2016). Promoting positive behaviour using the Good Behaviour Game: A meta-analysis of single-case research. *Journal of Positive Behaviour Interventions*, *18*, 180– 190.

Cohen, J. (2013). Statistical Power Analysis for the Behavioral Sciences. In *Statistical Power Analysis for the Behavioral Sciences*. Routledge. https://doi.org/10.4324/9780203771587

College Station TX: Texas A&M University. (n.d.). *NAP Calculator | Single Case Research*. Retrieved June 28, 2022, from http://singlecaseresearch.org/calculators/nap

Department for Edcuation. (2021). *Permanent exclusions and suspensions in England, Academic Year 2019/20 – Explore education statistics – GOV.UK*. https://explore-education-statistics.service.gov.uk/find-statistics/permanent-and-fixed-period-exclusions-in-england/2019-20

Ford, W. B., Radley, K. C., Tingstrom, D. H., & Dufrene, B. A. (2020). Efficacy of a No-Team Version of the Good Behavior Game in High School Classrooms. *Journal of Positive Behavior Interventions*, *22*(3), 181–190.

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.

Gresham, F. M., & Gresham, G. N. (1982). Interdependent, dependent, and independent group contingencies for controlling disruptive behaviour. *The Journal of Special Education*, *16*, 101–110.

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. https://doi.org/10.1002/jaba.513

Higgins, J., Williams, R., & McLaughlin, T. F. (2001). The effects of a token economy employing instructional consequences for a third-grade student with learning disabilities: A data-based case study. *Education & Treatment of Children*, *24*, 99–106.

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, A., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Council for Exceptional Children, 2*, 165–179.

Humphrey, N., Hennessey, A., Ashworth, E., Frearson, K., Black, L., Petersen, K., Wo, L., Panayiotou, M., Lendrum, A., Wigelsworth, M., Birchinall, L., Squires, G., & Pampaka, M. (2018). Good Behaviour Game: Evaluation report and executive summary. *Education Endowment Foundation.* https://educationendowmentfoundation.org.uk/public/files/GBG_evaluation_report.pdf.

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. /pmc/articles/PMC3188824/

Joslyn, P. R., Donaldson, J. M., Austin, J. L., & Vollmer, T. R. (2019). The Good Behavior Game: A brief review. In *Journal of Applied Behavior Analysis* (Vol. 52, Issue 3, pp. 811–815). Wiley-Blackwell Publishing Ltd. https://doi.org/10.1002/jaba.572

Law, M., Stewart, D., Pollock, N., Letts, L., Bosch, J., & Westmorland, M. (1998). *Guidelines for Critical Review* Form - Quantitative Studies. Retrieved from http://www-fhs.mcmaster.ca/rehab/ebp/http://wwwfhs.mcmaster.ca/rehab/ebp/pdf/qualguidelines.pdfhttp://wwwfhs.mcmaster.ca/rehab/ebp/pdf/qualreview.pdf

Little, S. G., Akin-Little, A., & O'Neill, K. (2015). Group contingency interventions with children—1980–2010: A metaanalysis. *Behavior Modification*, *39*, 322–341.

Mills, E. (2019). *How effective are technology-assisted, parent Naturalistic Developmental Behavioural Interventions for supporting the development of social-communication skills in young children with Autism Spectrum Disorder?* [UCL]. https://www.ucl.ac.uk/educational-psychology/resources/CS1Mills18-21.docx.pdf

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

Parker, R. I., & Vannest, K. (2009). An Improved Effect Size for Single-Case Research: Nonoverlap of All Pairs. *Behavior Therapy*, *40*(4), 357–367. https://doi.org/10.1016/j.beth.2008.10.006

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

Reinke, W., Herman, C., & Stormont, M. (2013). Classroomlevel positive behaviour supports in schools implementing SW-PBIS: Identifying areas for enhancement. *Journal of Positive Behavior Interventions*, *5*, 39–50.

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

Smith, S., Barajas, K., Ellis, B., Moore, C., McCauley, S., & Reichow, B. (2021). A Meta-Analytic Review of Randomized Controlled Trials of the Good Behavior Game. *Behavior Modification*, *45*(4), 641–666. https://doi.org/10.1177/0145445519878670

Stratton, K. K., Gadke, D. L., & Morton, R. C. (2019). Using the Good Behavior Game with High School Special Education Students: Comparing Student- and Teacher-Selected Reinforcers. *Journal of Applied School Psychology*, *35*(2), 105–121. https://doi.org/10.1080/15377903.2018.1509920

Tanol, G., Johnson, L., McComas, J., & Cote, E. (2010). Responding to rule violations or rule following: A comparison of two versions of the Good Behavior Game with kindergarten students. *Journal of School Psychology*, *48*, 337–355.

Troncoso, P., & Humphrey, N. (2021). Playing the long game: A multivariate multilevel non-linear growth curve model of long-term effects in a randomized trial of the Good Behavior Game. *Journal of School Psychology*, *88*, 68–84. https://doi.org/10.1016/j.jsp.2021.08.002

Vargo, K., & Brown, C. (2020). An evaluation of and preference for variations of the Good Behavior Game with students with autism. *Behavioral Interventions*, *35*(4), 560–570. https://doi.org/10.1002/bin.1740

Wilson, D. B. (n.d.). *Effect Size Calculator : The Campbell Collaboration - Framed Page*. Practical Meta-Analysis Effect Size Calculator. Retrieved June 28, 2022, from https://www.campbellcollaboration.org/research-resources/effect-size-calculator.html

Appendix A - Excluded studies

Table 1.

Studies excluded from the review after full-text screening based on exclusion criteria

	Full Reference	Exclusion Criteria number
1.	Bohan, C., Smyth, S., & McDowell, C. (2021). An Evaluation of the Caught Being Good Game with an Adolescent Student Population. *Journal of Positive Behavior Interventions*, *23*(1), 42–52. https://doi.org/10.1177/1098300720928455	1
2.	Ford, William Blake. (2017). Evaluation of a positive version of the Good Behavior Game utilizing ClassDojo technology in secondary classrooms. *ProQuest Dissertations and Theses*, 98. http://ergo.southwales.ac.uk/login?url=https://search.proquest.com/docview/1871769073?accountid=15324%0Ahttp://whel-primo.hosted.exlibrisgroup.com/openurl/44WHELF_USW/44WHELF_USW_services_page?	1
3.	Joslyn, P. R. (2017). Classroom management procedures with students who have histories of delinquency and emotional and behavioral disorders. In *Dissertation Abstracts International: Section B: The Sciences and Engineering* (Vol. 80, Issues 7-B(E)).	8 (a few databases have this article catalogued as 2019, which is why it was missed at the date screening phase, but its original date is 2017)
4.	Joslyn, P. R., Vollmer, T. R., & Kronfli, F. R. (2019). Interdependent Group Contingencies Reduce Disruption in Alternative High School Classrooms. *Journal of Behavioral Education*, *28*(4), 423–434. https://doi.org/10.1007/s10864-019-09321-0	1

Appendix B - Weight of Evidence A (WoE A)

To evaluate the methodological quality of each study, two protocols were used according to the research design utilised within the study. Judgements for four small-N or single case designs were evaluated using criteria from Horner et al., (2005) and for the one quantitative design, criteria from the Law et al. (1998) protocol was used. Both protocols had seven criteria each study was evaluated against, and the average provided the overall rating for WoE A.

Table 1a.

Summary of Weight of Evidence A: Ratings for Small N Designs

	Dimensions							
Authors	A. Participants and setting	B. Dependent variable	C. Independent variable	D. Baseline	E. Experimental control/ internal validity	F. External validity	G. Social validity	Overall WoE A
Ford et al. (2020)	3	3	3	2	3	2	2	2.57
Stratton et al. (2019)	1	3	3	2	3	3	2	2.43
Groves & Austin (2019)	2	3	2	2	3	3	2	2.43
Vargo & Brown (2020)	3	3	3	3	3	3	2	2.86

Table 1b.

Weight of Evidence A: Rating for Quantitative Studies

	Dimensions							
Author	1. Study Purpose	2. Literature	3. Design	4. Sample	5. Outcomes	6. Intervention	7. Results	Overall WoE A
Troncoso & Humphrey (2021)	3	3	2	3	3	2	2	2.57

Table 2.

Weight of Evidence A - Overall Rating descriptors for Small-N Design and Quantitative Studies

WoE A Rating	Criteria
High	Average rating across 7 judgement areas is 2.5 or above
Medium	Average rating across 7 judgement is between 1.5-2.4
Low	Average rating across 7 judgement areas is 1.4 or below

Table 3a.

Weight of Evidence A - Rating Criteria for Small-N Studies using Horner et al. (2005) as adapted by (Mills, 2019)

A. Description of Participants	- Participants are described with sufficient detail to allow others to select individuals with similar characteristics (e.g., age, gender, disability, diagnosis) - The process for selecting participants is described with replicable precision - Critical features of the physical setting are described with sufficient precision to allow replication
Rating	3 = All of the criteria are fulfilled 2 = Two of the criteria are fulfilled 1 = One of the criteria is fulfilled 0 = None of the criteria are fulfilled

B. Dependent Variable	- Dependent variables are described with operational precision - Each dependent variable is measured with a procedure that generates a quantifiable index - Measurement of the dependent variable is valid and described with replicable precision - Dependent variables are measured repeatedly over time - Data are collected on the reliability or interobserver agreement associated with each dependent variable, and IOA levels meet minimal standards
Rating	3 = All of the criteria are fulfilled 2 = Three or four of the criteria are fulfilled 1 = One or two of the criteria is fulfilled 0 = None of the criteria are fulfilled

C. Independent Variable	- Independent variable is described with replicable precision - Independent variable is systematically manipulated and under the control of the experimenter - There is overt measurement of the fidelity of the implementation for the independent variable
Rating	3 = All of the criteria are fulfilled 2 = Two of the criteria are fulfilled 1 = One of the criteria is fulfilled 0 = None of the criteria are fulfilled

D. Baseline	- The study includes a baseline phase that provides repeated measurement of the dependent variable(s) - The study establishes a pattern of responding that can be used to predict the pattern of future performance, if introduction or manipulation of the independent variable did not occur - Baseline conditions are described with replicable precision
Rating	3 = All of the criteria are fulfilled 2 = Two of the criteria are fulfilled 1 = One of the criteria is fulfilled 0 = None of the criteria are fulfilled

E. Experimental Control/Internal Validity	- The design provides at least three demonstrations of experimental effect at three different points in time - The design controls for common threats to internal validity (e.g., permits elimination of rival hypotheses) - The results document a pattern that demonstrates experimental control
Rating	3 = All of the criteria are fulfilled 2 = Two of the criteria are fulfilled 1 = One of the criteria is fulfilled 0 = None of the criteria are fulfilled

F. External Validity	- Experimental effects are replicated across participants and settings to establish external validity
Rating	3 = Experimental effects are replicated across 3 or more participants and in a unique setting 2 = Experimental effects are replicated across 3 or more participants 1 = Experimental effects are replicated across at least 2 participants 0 = Experimental effects are replicated with less than 2 participants

G. Social Validity	- The dependent variable is socially important - The magnitude of change in the dependent variable resulting from the intervention is socially important - Implementation of the independent variable is practical and cost effective - Social validity is enhanced by the implementation of the independent variable over extended time periods, by typical intervention agents, in typical physical and social contexts
Rating	3 = All of the criteria are fulfilled 2 = Two or three of the criteria are fulfilled 1 = One of the criteria are fulfilled 0 = None of the criteria are fulfilled

Table 3b.

Weight of Evidence A - Rating Criteria for Quantitative Studies from Law et al. (1998)

1. Study Purpose	Purpose of the study outlined Application to Educational Psychology stated Relevance to the research question of current review
Rating	3 = All of the criteria are fulfilled 2 = Two or three of the criteria are fulfilled 1 = One of the criteria are fulfilled 0 = None of the criteria are fulfilled

2. Literature	Review of the literature present providing background to the study Clinical importance stated Identifies gaps in current research and justifies the need for the study being reported
Rating	3 = All of the criteria are fulfilled 2 = Two or three of the criteria are fulfilled 1 = One of the criteria are fulfilled 0 = None of the criteria are fulfilled

3. Design	Study design described clearly Appropriateness of chosen study design for study question Consideration for biases that may influence results stated
Rating	3 = All of the criteria are fulfilled 2 = Two or three of the criteria are fulfilled 1 = One of the criteria are fulfilled 0 = None of the criteria are fulfilled

4. Sample	Detailed description of the participants with an indication of informed consent Comparisons between groups included demonstrating similarity of groups Sample size justification provided
Rating	3 = All of the criteria are fulfilled 2 = Two or three of the criteria are fulfilled 1 = One of the criteria are fulfilled 0 = None of the criteria are fulfilled

5. Outcomes	Clear outline of outcome measures with frequency of assessment recorded Reliability of measure reported Validity of measure reported
Rating	3 = All of the criteria are fulfilled 2 = Two or three of the criteria are fulfilled 1 = One of the criteria are fulfilled 0 = None of the criteria are fulfilled

6. Intervention	Detailed description of the intervention so that it could be replicated in practice Consideration shown for contamination bias consideration shown for avoiding cointervention
Rating	3 = All of the criteria are fulfilled 2 = Two or three of the criteria are fulfilled 1 = One of the criteria are fulfilled 0 = None of the criteria are fulfilled

7. Results	Results reported in terms of significance using an appropriate analysis method Clinical importance discussed Effect size reported
Rating	3 = All of the criteria are fulfilled 2 = Two or three of the criteria are fulfilled 1 = One of the criteria are fulfilled 0 = None of the criteria are fulfilled

Appendix C - Weight of Evidence B (WoE B)

The WoE B was evaluated through the use of criteria set out by Petticrew and Roberts' (2003) "Typology of Evidence", which is recommended as effective for answering questions about the effectiveness of an intervention according to the study design used. The ratings given to each study reviewed can be found in Table 1. Following that is an illustration of the criteria used, giving an indication of the type of rating that would be assigned to each study design (Table 2).

Table 1.

Weight of Evidence B – Ratings for each study reviewed

Authors	Overall WoE B
Ford et al., (2020)	2
Stratton et al. (2019)	2
Groves & Austin (2019)	2
Vargo & Brown (2020)	2
Troncoso & Humphrey (2021)	3

Table 2.

Weight of Evidence B - Criteria for Ratings

WoE B Rating	Study Design	Further Criterion
3	Randomised control trials Randomised experimental	Pre and post collection of data for all groups & Minimum of one control and comparison group
2	Quasi-experimental design	Pre and post collection of data for all groups & Minimum of one control and comparison group
	Small N/ single case design	Single/ small N designs should have a minimum of 3 experimental effects occasions displayed (across 3 participants or 3 varying time points within 1 participant)
	Cohort Studies	
1	Non-experimental study designs	Pre and post collection of data for all groups
	Qualitative research	No control and comparison group
	Other Small N designs & Surveys	For single N designs there is less than 3 occasions where experimental effect is displayed

These criteria are informed by "Typology of Evidence" recommendations for research most suitable to examine the effectiveness of interventions (Petticrew & Roberts, 2003)

Appendix D - Weight of Evidence C (WoE C)

WoE C seeks to appraise how relevant each of the reviewed studies were at answering how effective the Good Behaviour Game was at reducing displays of disruptive behaviour in secondary classrooms settings. The studies were rated according to their appropriateness towards answering the review question and they were given a rating from 1-3, based on three criteria (Table 2) upon which judgements were made. These ratings make up WoE C (Table 1).

Table 1.

Weight of Evidence C – Ratings

Authors	Participant characteristics	Setting	Variations of the intervention	Overall WoE C
Ford et al., (2020)	3	2	2	2.33
Stratton et al. (2019)	2	2	3	2.33
Groves & Austin (2019)	3	3	2	2.67
Vargo & Brown (2020)	3	2	2	2.33
Troncoso & Humphrey (2021)	2	3	3	2.67

Table 2.

Weight of Evidence C - Criteria for Ratings

Criteria	WoE Rating	Descriptor	Rationale
Participant characteristics	3	High level of disruptive behaviour displayed by pupils	
	2	Pupils do not display high levels of disruptive behaviour	Intervention is most effective for pupils displaying disruptive behaviours in classrooms Stratton et al. (2019).
	1	No reference to display of disruptive behaviour made	
Setting	3	Schools within the United Kingdom	Research conducted within the UK or countries of similar standing would increase the generalisability of the

	2	Schools within OECD countries	findings to the British context and this will be of greater relevance to Educational Psychology practice within UK schools.
	1	Schools outside of OECD countries	
Variations of the intervention	3	The intervention is based on the original version of GBG	
	2	The intervention is an adapted or enhanced version of GBG with an explanation of how it differs from the original	The researcher is interested in how the main components of the GBG in its original form impacted behaviour modification outcomes.
	1	The intervention is an adapted or enhanced version of GBG without an explanation of how it differs from the original	

Appendix E – Coding Protocols for Small-N Studies (Single Case Designs)

Coding Protocol for Small N Designs – Ford et al., 2020

[Adapted from 'The Use of Single Subject Research to Identify Evidence - Based Practice Special Education' Horner, Carr, Halle, McGee & Wolery (2005)]

Study Reference: Ford, W. B., Radley, K. C., Tingstrom, D. H., & Dufrene, B. A. (2020). Efficacy of a No-Team Version of the Good Behavior Game in High School Classrooms. *Journal of Positive Behavior Interventions*, *22*(3), 181–190. https://doi.org/10.1177/1098300719890059

Type of Publication: ☐Book/Monograph ☒Journal Article ☐Book Chapter ☐Other (specify):	**Study Type:** - Single case A/B/A/B withdrawal design (Baseline, treatment, treatment withdrawn, reimplementation of treatment). **Intervention name and description:** - Good Behaviour Game (GBG) no-teams version
A. Description of Participants **Rating:** **☒3 ☐2 ☐1 ☐0**	A1. Participants are described with sufficient detail to allow others to select individuals with similar characteristics (e.g., age, gender, disability, diagnosis). ☒Yes ☐No A2. The process for selecting participants is described with replicable precision. ☒Yes ☐No A3. Critical features of the physical setting are described with sufficient precision to allow replication. ☒Yes ☐No
B. Dependent Variable	B1. The dependent variables are described with operational precision. ☒Yes ☐No B2. Each dependent variable is measured with a procedure that generates a quantifiable index. ☒Yes ☐No B3. Measurement of the dependent variable is valid and described with replicable precision. ☒Yes ☐No B4. The dependent variables are measured repeatedly over time. ☒Yes ☐No B5. Data are collected on reliability or interobserver agreement

Rating: ☒**3** ☐**2** ☐**1** ☐**0**	associated with each dependent variable and IOA levels meet minimal standards (e.g., IOA = 80%; Kappa = 60%). ☒Yes ☐No
C. Independent Variable **Rating:** ☒**3** ☐**2** ☐**1** ☐**0**	C1. The independent variable is described with replicable precision ☒Yes ☐No C2. The independent variable is systemically manipulated and under the control of the experimenter. ☒Yes ☐No C3. There is overt measurement of the fidelity of the implementation of the independent variable. ☒Yes ☐No
D. Baseline **Rating:** ☐**3** ☒**2** ☐**1** ☐**0**	D1. The study includes a baseline phase that provides repeated measurement of the dependent variable(s). ☒Yes ☐No D2. The study establishes a pattern of responding that can be used to predict the pattern of future performance if introduction or manipulation of the independent variable didn't occur. ☐Yes ☒No Baseline conditions are described with replicable precision. ☒Yes ☐No
E. Experimental Control/Internal Validity **Rating:** ☒**3** ☐**2** ☐**1** ☐**0**	E1. The design provides at least three demonstrations of experimental effect at three different points in time. ☒Yes ☐No E2. The design controls for common threats to internal validity e.g., permits elimination of rival hypotheses. ☒Yes ☐No E3. The results document a pattern that demonstrates experimental control ☒Yes ☐No
F. External Validity **Rating:** ☐**3** ☒**2** ☐**1** ☐**0**	F1. Experimental effects (select one) ☐Experimental effects are replicated across three or more participants and in a unique setting ☒Experimental effect are replicated across three or more participants ☐Experimental effect are replicated across at least two participants ☐Experimental effects are replicated with less than 2 participants

G. Social Validity **Rating:** ☐**3** ☒**2** ☐**1** ☐**0**	G1. The dependent variable is socially important ☒Yes ☐No G2. The magnitude of change in the dependent variable resulting from the intervention is socially important. ☐Yes ☒No G3. Implementation of the independent variable is practical and cost effective. ☐Yes ☒No G4. Social validity is enhanced by implementation of the independent variable over extended time periods, by typical intervention agents, in typical physical and social contexts. ☒Yes ☐No

Dimension	Rating
A. Description of participants	3
B. Dependent Variable	3
C. Independent Variable	3
D. Baseline	2
E. Experimental Control/Internal Validity	3
F. External Validity	2
G. Social Validity	2
Total	18
Overall WoE A	18/7 = 2.57

Appendix F – Coding Protocol for Quantitative Studies

Critical Review Form - Quantitative Studies

Study Reference: Troncoso, P., & Humphrey, N. (2021). Playing the long game: A multivariate multilevel non-linear growth curve model of long-term effects in a randomized trial of the Good Behavior Game. *Journal of School Psychology*, *88*, 68–84. https://doi.org/10.1016/j.jsp.2021.08.002	
JUDGEMENTS AND CONSIDERATIONS:	
1. STUDY PURPOSE: **Rating:** ☒3 ☐2 ☐1 ☐0	*Outline the purpose of the study. How does the study apply to Educational Psychology and/or your research question?* ➢ Purpose of the study outlined - Yes This study examined the impact of the Good Behaviour Game (GBG) on children's developmental trajectories of disruptive behaviour, concentration problems, and prosocial behaviour from middle childhood (ages 6–7 years) to early adolescence (ages 10–11 years). Application to Educational Psychology stated - Yes Relevance to the research question of current review - Yes
2. LITERATURE **Rating:** ☒3 ☐2 ☐1 ☐0	*Describe the justification of the need for this study.* Review of the literature present providing background to the study - Yes Clinical importance stated – Yes Especially relevant for the context of the UK (study completed in England). Identifies gaps in current research and justifies the need for the study being reported – Yes growth curve model used, as opposed to point-in-time scope usually completed in with this type of study; long term effects of the intervention was examined long-term; intervention effects were considered at both main and subgroup levels - differences examined across a variety of domains including sex and conduct problems risk status, with effects examined separately and in

	combination. Intended contribution extends the knowledge base regarding the scope, specificity, and timing of intervention effect.
3. DESIGN: • Randomized (RCT) • Cohort • Single Case Design • Before and after • Case-control • Cross-sectional • Case Study **Rating:** ☐ **3** ☒**2** ☐**1** ☐**0**	*Describe the study design. Was the design appropriate for the study question? (e.g., for knowledge level about this issue, outcomes, ethical issues, etc.)* Study design described clearly – Yes, Multivariate multilevel non-linear growth curve model in a cluster randomized controlled trial (RCT) - Seventy-seven schools were randomly allocated by an independent trial unit to deliver the GBG (intervention) or continue usual practice (control) for a period of two years. - A minimization algorithm was used to ensure balance across trial arms with respect to school size and the proportion of children eligible for free school meals. - Outcome data were collected at baseline (pre-randomization, Time 1 [T1]) and then annually on four further occasions (Time 2 [T2], Time 3 [T3], Time 4 [T4], Time 5 [T5]). - T1 to T3 represents the period of GBG implementation in the intervention arm of the trial - T3 to T5 represents a clean follow-up phase i.e., none of the trial sample were exposed to the GBG during this period. Appropriateness of chosen study design for study question – Yes – partially *Specify any biases that may have been operating and the direction of their influence on the results.* Consideration for biases that may influence results stated - No
4. SAMPLE: N = 3084 **Rating:** ☒**3** ☐**2** ☐**1** ☐**0**	*Sampling (who; characteristics; how many; how was sampling done?) If more than one group, was there similarity between the groups?* Detailed description of the participants with comparisons between groups included demonstrating similarity of groups. – Yes, extensively described. Intervention group – N = 1497 Control group – N = 1469 Sample size justification provided – Yes *Describe ethics procedures. Was informed consent obtained?*

<table>
<tr><td></td><td>Indication of informed consent provided – Yes and ethical approval was granted, as well as opt-out consent was sought.</td></tr>
<tr><td>5. OUTCOMES:

Rating: ☒3 ☐2 ☐1 ☐0</td><td>Specify the frequency of outcome measurement (i.e., pre, post, follow-up)
<table><tr><td>Outcome areas:
• Concentration problems
• Disruptive behaviour
• Prosocial behaviour</td><td>List measures used:
• Teacher Observation of Child Adaptation Checklist structured (TOCA-C)
• Observation Schedule
• Strength and Difficulties Questionnaire (SDQ</td></tr></table>
Reliability of measure reported – Yes

Validity of measure reported – Yes</td></tr>
<tr><td>6. INTERVENTION:

Rating: ☐3 ☒2 ☐1 ☐0</td><td>Provide a short description of the intervention (focus, who delivered it, how often, setting). Could the intervention be replicated in Educational Psychology practice?

Intervention was described in detail?

Contamination was avoided – Yes
Cointervention was avoided – No</td></tr>
<tr><td>7. RESULTS:</td><td>What were the results? Were they statistically significant (i.e., p < 0.05)? If not statistically significant, was study big enough to show an important difference if it should occur? If there were multiple outcomes, was that considered for the statistical analysis?

Results reported in terms of significance using an appropriate analysis method – Yes, extensively
• concentration problems and prosocial behaviour decreased each year, with standard deviations remaining relatively stable.
• disruptive behaviour as observed means tended to increase with time.</td></tr>
</table>

Rating: ☐ 3 ☒2 ☐1 ☐0	• Analysis revealed that the intervention altered trajectories of concentration problems, with those exposed to the GBG experiencing a mean linear decrease of 0.151 SD with respect to the previous year (and strong support for better outcomes at T5), relative to their counterparts in control schools. • did not find reliable evidence of an intervention effect on trajectories of disruptive behaviour or prosocial behaviour *What was the clinical importance of the results? Were differences between groups clinically meaningful? (If applicable)* Clinical importance discussed – Yes robust evidence that the GBG influences the trajectory of children's concentration problems over time. The direction of this effect is consistent with both the theorized effects of the intervention and developmental trends in children's capacity to pay attention, stay on task, and resist distractions during the elementary school years. It also provides evidence for the long-term effects of the game in the secondary school follow-up period. Effect size was reported? No
Drop-outs were reported?	*Did any participants drop out from the study? Why? (Were reasons given and were drop-outs handled appropriately?)* Yes – partially explained
CONCLUSIONS AND CLINICAL IMPLICATIONS: Conclusions were appropriate given study methods and results	*What did the study conclude? What are the implications of these results for Educational Psychology practice? What were the main limitations or biases in the study?* This study demonstrated the impact of the Good Behaviour Game on children's developmental trajectories of concentration problems, in addition to resulting in notable improvements in prosocial behaviour among those with elevated conduct problems. In doing so, it highlighted the value and utility of growth curve modelling of intervention effects and including data points that extend well beyond the conclusion of a given period of implementation. In some ways demonstrated that playing the "long game" may come with benefits. The study also showed that the GBG had no effects in relation to disruptive behaviour.
Sum of all scored dimensions - 3+3+2+3+3+2+2 = **18** **WEIGHTING** (sum of all dimensions divided by the number of dimensions) – 18/7 = **2.57**	