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

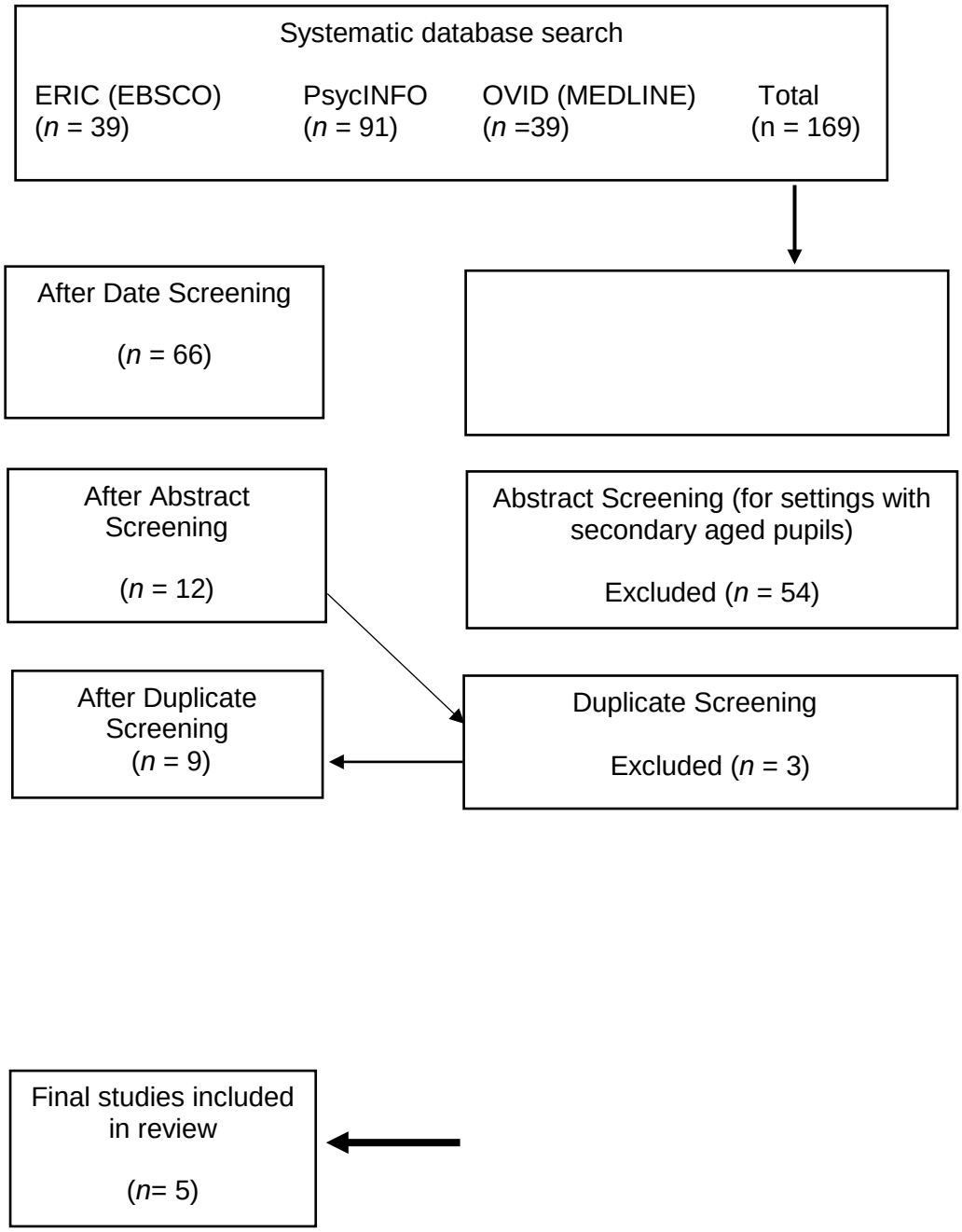
(Skinner, 1945). This theory highlights that desired behaviours can be

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

Figure 1.

Flow Chart Illustrating the Process of the Systematic Literature Search



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., TinE.002 Tc 10.42.5o4 67.7,u5ey. Cte9 >>BDC

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

	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)

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)

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

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

Measures

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

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

			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

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

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.

Mills, E. (2019). *How effective are technology-assisted, parent Naturalistic*

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

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)

[redacted]

t review

	Detailed description of the participants with an indication of informed consent

4. Sample

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

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,1

Appendix F – Coding Protocol for 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:

Outline the purpose of the study. How does the study apply to Educational Psychology and/or your research question?

¾ Purpose of the study outlined - Yes

This study examined the impact of the Good
Behe ehe hi Tw 0.7>>BDCi3 (y)16.1903 / (y)16.190 429.2 /P <</MCID 52BT

Rating: **• 3 • 2 • 1 • 0**

combination. Intended contribution extends the knowledge base regarding the scope, specificity, and timing of intervention effect.

	Indication of informed consent provided – Yes and ethical approval was granted, as well as opt-out consent was sought.

Specify the frequency of outcome measurement (i.e., pre, post, follow-up)

5. OUTCOMES:

Outcome areas:

- x Concentration problems
- x Disruptive behaviour
- x Prosocial behaviourm

n o i i v a h e b

Rating: • 3 • 2 • 1 • 0

Reliability of measure reported – Yes

Validity of measure reported – Yes