



Education  
Endowment  
Foundation

# Good Behaviour Game

Evaluation report and executive summary

July 2018

## **Independent evaluators:**

Neil Humphrey, Alexandra Hennessey, Emma Ashworth, Kirsty Frearson, Louise Black, Kim Petersen, Lawrence Wo, Margarita Panayiotou, Ann Lendrum, Michael Wigelsworth, Liz Birchinall, Garry Squires and Maria Pampaka.

MANCHESTER  
1824

The University of Manchester



The Education Endowment Foundation (EEF) is an independent grant-making charity dedicated to breaking the link between family income and educational achievement, ensuring that children from all backgrounds can fulfil their potential and make the most of their talents.

The EEF aims to raise the attainment of children facing disadvantage by:

- identifying promising educational innovations that address the needs of disadvantaged children in primary and secondary schools in England;
- evaluating these innovations to extend and secure the evidence on what works and can be made to work at scale; and
- encouraging schools, government, charities, and others to apply evidence and adopt innovations found to be effective.

The EEF was established in 2011 by the Sutton Trust as lead charity in partnership with Impetus Trust (now part of Impetus - Private Equity Foundation) and received a founding £125m grant from the Department for Education.

Together, the EEF and Sutton Trust are the government-designated What Works Centre for improving education outcomes for school-aged children.



**For more information about the EEF or this report please contact:**

**Danielle Mason**  
Head of Research

Education Endowment Foundation  
9th Floor, Millbank Tower  
21–24 Millbank  
SW1P 4QP  
p: 020 7802 1679  
e: [danielle.mason@eefoundation.org.uk](mailto:danielle.mason@eefoundation.org.uk)  
w: [www.educationendowmentfoundation.org.uk](http://www.educationendowmentfoundation.org.uk)

## About the evaluator

The project was independently evaluated by a team from the Manchester Institute of Education, University of Manchester. The evaluation team were: Neil Humphrey, Alexandra Hennessey, Emma Ashworth, Kirsty Frearson, Louise Black, Kim Petersen, Lawrence Wo, Margarita Panayiotou, Ann Lendrum, Michael Wigelsworth, Liz Birchinal, Garry Squires and Maria Pampaka.

The lead evaluator was Professor Neil Humphrey.

### **Contact details:**

Professor Neil Humphrey

Manchester Institute of Education

The University of Manchester

Oxford Road

Manchester

M13 9PL

**p:** 0161 275 3404

**e:** [neil.humphrey@manchester.ac.uk](mailto:neil.humphrey@manchester.ac.uk)

## Contents

About the evaluator.....	2
Contents.....	3
Executive summary.....	5
Introduction .....	7
Methods .....	19
Impact evaluation .....	31
Implementation and process evaluation .....	40
Conclusion.....	67
References.....	73
Appendix 1: Memorandum of agreement.....	81
Appendix 2: Parent information sheet.....	87
Appendix 3: Parent consent form .....	93
Appendix 4: Confirmatory factor analyses for teacher surveys .....	95
Appendix 5: GBG observation schedule.....	100
Appendix 6: Exploratory factor analyses .....	101
Appendix 7: Single sample t-tests comparing trial school characteristics ....	102
Appendix 8: Predictors of missing data at follow-up.....	103
Appendix 9: MLM ITT and sub-group analyses .....	104
Appendix 10: Analyses of teacher-level outcomes .....	112
Appendix 11: Implementation analyses .....	113
Appendix 12: Security classification of trial findings .....	121
Appendix 13: EEF cost rating .....	122



## Executive summary

### The project

The aim of the Good Behaviour Game (GBG) is to improve pupil behaviour through the implementation of a behaviour management system with the following core elements: classroom rules, team membership, monitoring of behaviour, and positive reinforcement (rewards). It is a universal intervention and is therefore delivered to all children in a given class by their teacher. Over the course of implementation, it is intended that there is a natural progression in terms of the types of rewards given (from tangible rewards such as stickers to more abstract rewards such as free time), how long the game is played for (from 10 minutes to a whole lesson), at what frequency (from three times a week to every day), and when rewards are given (at the end of the game, the end of the day, and the end of the week). Teachers receive two days of initial training, with one day of follow-up training midway through the first year of implementation. On-going support for implementation is provided by trained GBG coaches employed by the delivery organisation, Mentor UK (who were in turn supported by the American Institutes for Research for this trial).

We used a randomised controlled trial design in which 77 schools were randomly allocated to implement the GBG for two years (38 schools) or continue their normal practices (39 schools). The target cohort was pupils in Year 3 (aged 7-8) in the first year of implementation (N=3084). The project was designed as an efficacy trial. Alongside the assessment of outcomes, we undertook a comprehensive mixed-methods implementation and process evaluation involving observations, interviews and focus groups. Delivery started in September 2015 and concluded in July 2017.

### Key conclusions

1. We found no evidence that the GBG improves pupils' reading. This result has a high security rating.
2. We found no evidence that the GBG improves pupils' behaviour (specifically, concentration problems, disruptive behaviour, and pro-social behaviour).
3. Implementation was variable and in particular, the frequency and duration with which the GBG was played did not reach the levels expected by the developer. One-quarter of schools in the intervention arm ceased implementation before the end of the trial.
4. Higher levels of pupil engagement with the game were associated with improved reading, concentration, and disruptive behaviour scores at follow-up. There was no clear evidence that other aspects of implementation (for example, how well or how frequently the game was played) were related to whether pupil outcomes improved. These results were sensitive to changes in how we analysed the data, and so should be interpreted with caution.
5. There was tentative evidence that boys identified as at-risk of developing conduct problems at the beginning of the project benefitted from the GBG. For these children, small reductions in concentration problems and disruptive behaviour were observed.

### EEF security rating

These findings have a high security rating (see appendix 11). This was an efficacy trial, which tested whether the intervention worked under developer-led conditions in a number of schools. It was a well-designed, two-armed randomised controlled trial. The study was well-powered and the pupils in GBG schools were similar to those in the comparison schools in terms of prior attainment. However, the following factor reduced the security of the trial: 19% of the pupils who started the trial were not included in the final analysis, because they had moved school or were absent on the day of testing.

## Additional findings?

Our analyses indicated that the GBG had no significant impact on pupils' reading, concentration problems, disruptive behaviour, or pro-social behaviour when compared to those attending comparison schools. There was tentative evidence that boys at-risk of developing conduct problems benefitted from the GBG in relation to their concentration problems and disruptive behaviour. However, there was no evidence of similar differential gains among pupils eligible for Free School Meals (FSM). Implementation was variable and in particular, the frequency and duration with which the GBG was played did not reach the levels expected by the developer. Furthermore, one-quarter of schools in the intervention arm ceased implementation before the end of the trial. Teachers cited several reasons for discontinuing the GBG. For some it was a problem of utility: the game took time and effort to set up that was not outweighed, in their view, by the benefits of playing. Some teachers felt the game did not fit with all curriculum content, and so competed with valuable classroom activities rather than complementing them. For others, the strict rule that the teacher could not interact with students during the game was seen to impede the extent to which they could aid their academic progress. This was seen as a particular problem where students had additional needs.


Higher levels of participant responsiveness (the extent to which children engaged with the GBG) were associated with significantly improved reading, concentration problems, and disruptive behaviour scores at follow-up. Finally, there was no significant impact of the intervention on teacher stress, self-efficacy in classroom management, or retention.

## Cost?

The estimated initial start-up cost per school is £4,000, **£37.04 per pupil**. Over three years there would be some savings (e.g. initial teacher training), such that in subsequent years the cost per school would average £3,500. Over 3 years, the cost would average **£33.95 per pupil**. If the number of schools buying the programme were higher, then costs would be lower. For example, Mentor UK have also budgeted for a 60-school programme with initial start-up cost of £3,700 per school, £35.53 per pupil.

As the GBG is played during a typical lesson/activity, there is minimal additional teaching time or staffing requirements outside of normal practice. Preparation time is estimated to be marginal and would typically involve the organisation of pupils into teams, allocation of team leaders, and maintaining the posters and resources. There are also monthly GBG coach visits, which typically involve an observation of the game followed by a meeting of up to 30 minutes for discussion.

## Summary of impact on primary outcome—reading test scores

Outcome/ Group	Effect size (95% Confidence Interval)	Estimated months' progress	EEF security rating	No. of pupils	P value	EEF cost rating
<b>Reading</b>	0.03 (-0.08 to 0.16)	0		2504	0.30	£ £ £ £ £
<b>Reading – FSM pupils</b>	0.05 (-0.07 to 0.18)	1	N/A	591	0.22	£ £ £ £ £

## Introduction

### Intervention

The Good Behaviour Game (hereafter referred to as the GBG) is a universal behaviour management intervention. While it is primarily used with children in primary schools, it can also be implemented in early years and secondary education settings (Flower, McKenna, Bunuan, Muething, & Vega, 2014; Tingstrom, Sterling-Turner, & Wilczynski, 2006). The GBG was originally developed in the United States nearly 50 years ago (Barrish, Saunders, & Wolf, 1969), and since then versions of it have been utilised across a range of countries, to cater for culturally, linguistically and socio-economically diverse student populations (Nolan, Houlihan, Wanzek, & Jenson, 2014), including the United Kingdom (UK; Coombes, Chan, Allen, & Foxcroft, 2016), Sudan (Saigh & Umar, 1983), Belize (Nolan, Filter, & Houlihan, 2014), Belgium (Leflot, van Lier, Onghena, & Colpin, 2013), the Netherlands (Dijkman, Harting, & van der Wal, 2015), Spain (Ruiz-Olivares, Pino, & Herruzo, 2010), and Chile (Pérez, Rodríguez, De la Barra, & Fernández, 2005). It is included in the National Registry of Evidence-Based Programmes and Practices (<http://nrepp.samhsa.gov/landing.aspx>) and the Early Intervention Foundation's Guidebook (<http://guidebook.eif.org.uk/>), and is rated as a 'promising programme' in the Blueprints for Healthy Youth Development database (<http://blueprintsprograms.com/>).

In order to provide a comprehensive and transparent description of the GBG, we utilise an adapted version of the Template for Intervention Description and Replication (TIDieR) (Hoffmann et al., 2014), as per recommended reporting guidance (Humphrey et al., 2016), alongside a logic model to illustrate the theorised processes by which the intervention inputs lead to specified outcomes (Figure 1, below):

#### 1. Brief name

The Good Behaviour Game (GBG)

#### 2. Why (rationale/theory)

The GBG is underpinned by three key theories pertaining to human development: behaviourism (specifically, contingency management; Skinner, 1945), social learning theory (Bandura, 1986), and life course/social field theory (LCSFT; Kellam et al., 2011). In terms of behaviourism, a key assumption of the intervention is that behaviours that are rewarded are more likely to be reproduced. Thus, in the GBG, children receive positive reinforcement when they engage in appropriate behaviours (e.g. following the teacher's instructions during an activity). However, the group-based orientation of the intervention means it also draws upon social learning theory, in that children at-risk of developing conduct problems are able to learn from the appropriate behaviour being modelled effectively by other team members. Finally, LCSFT posits that successful adaptation at different life stages is contingent upon an individual's ability to meet particular social task demands. In school, these task demands include being able to pay attention, work well with others, and obey rules. Success in social adaptation is rated both formally and informally by other members of the social field (e.g. teachers, peers). LCSFT predicts that improving the way in which teachers socialise children (for example, adopting a more explicit approach to highlighting and promoting social task demands, as in the GBG) will improve their social adaptation. It is also predicted that early improvements in social adaptation in the classroom will lead to better adaptation to other social fields (e.g. peer group, family, work) later in life (Kellam et al., 2011).

#### 3. Who (recipients)

The GBG is a universal intervention and is therefore delivered to all children in a given class.

#### 4. What (materials)



Participating schools receive GBG manuals that detail the programme theory, goals and procedures. Other materials include some tangible rewards (e.g. stickers), displays (e.g. scoreboard, rules posters), and data forms for recording and monitoring purposes. In the current study, two additional resources were developed by a member of the evaluation team (Wo) following a request from the delivery team (Mentor UK). First, an online GBG scoreboard was created. Each teacher was able to log into a secure website to record games and probe data (see section 5 below) in real time and retrospectively, which could then be downloaded to assess temporal trends and inform future implementation planning. In turn, each GBG coach (see section 11 below) was able to access their assigned teachers' data for use in later support sessions, and the research team were able to access all teachers' data so that it could be used to monitor the length and frequency of games (see the implementation and process evaluation, IPE). Second, an electronic version of the fidelity checklist used by GBG coaches was developed. This was identical to the paper version used by the licensing organization (American Institutes for Research; AIR) and was used for the same purpose (e.g. to facilitate feedback following an observation session).

### 5. What (procedures)

The GBG is described by Tingstrom et al. (2006) as an "interdependent group-oriented contingency management procedure" (p. 225). The teacher divides the class into mixed teams with up to 7 members<sup>1</sup>. Where possible, each team should be balanced with equal representation of salient factors such as behaviour, academic ability, and gender. The teams then attempt to win the game as a means to access particular privileges/rewards. The game is played during a typical class activity. During the game period, the class teacher records the number of infractions to the following four rules among the teams:

- (1) We will work quietly
- (2) We will be polite to others
- (3) We will get out of seats with permission, and
- (4) We will follow directions.

In relation to the first rule, adherence is defined as working at a noise level that is deemed to be *appropriate* for the classroom activity being undertaken while the GBG is being played. Prior to the commencement of the game, the teacher agrees one of the following noise levels with the class: Level 0 (Voices Off, silence), Level 1 (Whisper, only the person sat next to you can hear you), Level 2 (Inside Voice, only people sat at your table can hear you), Level 3 (Speaker, your classmates can hear you), and Level 4 (Outside, 'playground' voice).

The game is 'won' by the team(s) with four or fewer infractions, who then access an agreed reward (Chan, Foxcroft, Smurthwaite, Coombes, & Allen, 2012; Kellam et al., 2011). The procedures undertaken before, during and immediately after a game session are detailed in the aforementioned intervention manual, as follows:

#### *Before game:*

- Teacher explains the task/activity
- Teacher checks understanding of the task/activity
- Teacher reminds pupils that they cannot ask for help
- Pupils are in teams of between 3 and 7 (except for special circumstances)<sup>2</sup>

---

<sup>1</sup> Team membership is typically varied several times in a school year (e.g. every half term).

<sup>2</sup> This might include, for example, a situation in which a child is placed in a team on their own as a response to them deliberately and repeatedly sabotaging their team's efforts to win the game.

- Pupils are in clear teams
- Teams are gender balanced
- Rules are appropriately verbally reviewed with the class
- Exemplars are modelled/described by the teacher and/or pupils
- Infractions are modelled/described by the teacher
- Infractions are described, but not modelled, by students
- Voice level for the task/activity is given by the teacher
- Teacher states when the game begins
- Teacher states how long the game will be played for
- Teacher sets a timer
- Teacher states that they will monitor infractions
- Teacher states that 4 infractions are permitted per team
- Teacher reminds pupils that they are not competing against each other

*During game:*

- Teacher records infractions on the scoreboard
- Teacher identifies infractions when they occur
- Teacher identifies rule breaking team (e.g. “Team 4 have broken rule 4, ‘we will follow directions’”)
- Teacher discreetly indicates infraction to specific pupil
- Rest of team and/or class praised for adhering to rules (e.g. “Well done everyone else for following rule 4”)
- Teacher does not punish pupils/teams for infractions
- Teacher monitors behaviour
- Teacher does not interact with pupils
- Teacher adheres to time limit set
- Teacher announces the end of the game

*After game:*

- Teacher repeats 4 infractions or less criterion
- Teacher announces winning team(s) only
- Members of winning team receive stamp (or marker etc.) in individual booklets
- Star placed on wall chart (or equivalent)

Over the course of implementation of the GBG, it is intended that there is a natural progression in terms of the types of rewards used (from tangible rewards such as stickers to more abstract rewards such as free time), how long the game is played for (from 10 minutes to a whole lesson), at what frequency (from three times a week to every day), and when rewards are given (at the end of the game, end of the day, and at the end of the week) (Elswick & Casey, 2011; Tingstrom et al., 2006). This progression is designed to maintain responsiveness, interest and challenge for students, while also encouraging generalisation. Thus good behaviour achieved during the relatively brief ‘game’ periods is increasingly generalised to other activities and parts of the school day. The intervention aims to build intrinsic reinforcement so that modified behaviour is retained even after external reinforcement is removed (maintenance) and will be exhibited in all settings (generalisation). These processes are documented by ‘game’ and ‘probe’ data collected by teachers during implementation (Chan et al., 2012). Probe data, used to assess generalisation, are collected *covertly* during an ordinary task/activity following the same procedures as in a game session (such as the teacher monitoring rule infractions among teams) but without explicitly setting up the rules and announcing infractions.

## **6. Who (provider)**

The GBG is implemented by class teachers.

### **7. How**

The GBG is implemented face-to-face during the normal school day. As it is a behaviour management strategy rather than a taught curriculum, it does not require an explicit 'space' in the class timetable, thereby minimising the displacement of other activities. However, the pre- and post-game procedures undertaken by the teacher (e.g., reminding the class of the rules, announcing the winners and providing rewards – see section 5 above) mean that some time is taken up before and after the game period/class activity.

### **8. Where**

The GBG is implemented on-site in participating schools.

### **9. When and how much**

The GBG is played throughout the school year. As noted above, dosage evolves throughout the period of implementation in terms of both the duration of the game (from 10 minutes to a whole lesson), and the frequency at which it is played (from three times a week to every day).

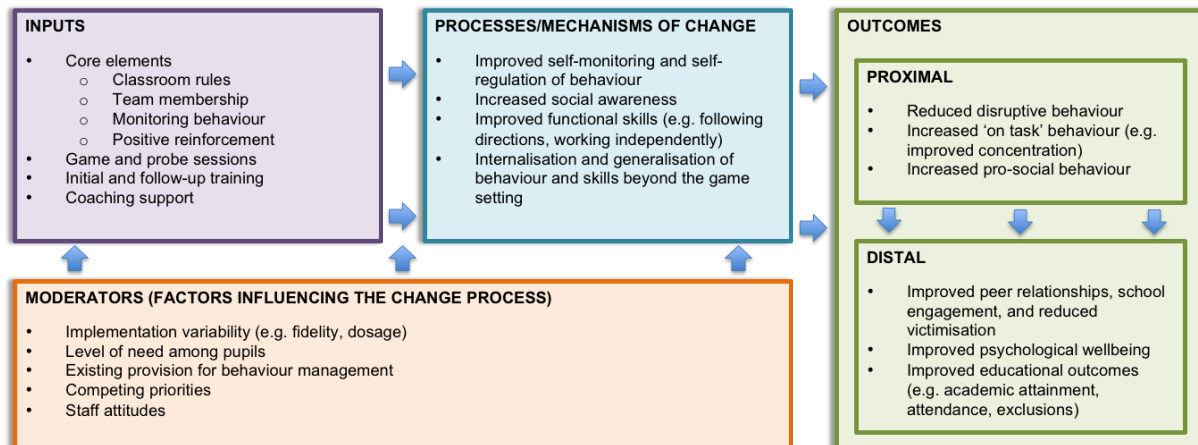
### **10. Tailoring**

The GBG is a manualised intervention and participating teachers receive initial and follow-up training in addition to technical support and assistance as a means to optimise fidelity of implementation. However, it is now accepted in the field of implementation science that some form of adaptation is inevitable and indeed may be desirable in order to improve local ownership and fit to context (Durlak & DuPre, 2008; US Department of Health and Human Services, 2002). An important aspect of the GBG coach role is to support teachers to make adaptations that are in keeping with the goals and theory of the intervention (Moore, Bumbarger, & Cooper, 2013).

### **11. How well (planned)**

Teachers receive three days of training (two days initial; one day follow-up) by coaches (mostly former teachers) contracted by Mentor, and trained by AIR. On-going technical support and assistance is provided by these trained coaches. In the current study, participating schools were each allocated a GBG coach who paid approximately monthly visits to support their implementation throughout the trial. These visits typically comprised modelling of game sessions, observation and feedback (including review of game, probe and fidelity checklist data – see above), ad-hoc email and telephone support, and provision of additional/booster training or information sessions as required.

Figure 1. Logic model for the GBG.



Note that the TIDieR framework and logic model information above represents the specific version of the GBG held under licence by AIR. The other most widely used version, the Pax Good Behaviour Game, follows a similar model to that outlined above, but differs in respect to (a) the language used to describe rule abidance and infractions (referred to as 'Pax' and 'spleems' respectively); (b) the game threshold (teams with 3 or fewer spleems as opposed to 4 or fewer infractions access the agreed reward); (c) parent activities to promote generalisation of self-regulation skills to the home environment and, (d) various additional idiosyncratic procedures (e.g. 'Pax Stix': random selection of students for potential reinforcement; PAX Quiet: hand signals used by the teacher; Tootles: teacher-written praise notes) (Weis, Osborne, & Dean, 2015).

Given the length of time since its inception and its subsequent widespread use across different countries and cultures, further variations of the GBG have proliferated, including those exploring its application in different phases of education (e.g. preschool, <5 years, Swiezy, Matson, & Box, 1993; high school, up to 18 years, Lynne, 2015), across a range of settings extending beyond standard classroom activities (e.g. school cafeteria, McCurdy, Lannie, & Barnabas, 2009; after school clubs, Philips Smith et al., 2014; school library, Fishbein & Wasik, 1981), and outside of mainstream education (e.g. in a special school for children with psychiatric disorders, Breeman et al., 2016) incorporating adaptations/modifications to intervention procedures (e.g. monitoring rule following as opposed to infractions, Tanol, Johnson, McComas, & Cote, 2010; adding the opportunity to win 'bonus points' for meeting classroom behaviour goals to offset losses earned by rule infraction, McGoey, Schneider, Rezzetano, Prodan, & Tankersley, 2010) and in combination with other interventions (e.g. Say-Do-Report correspondence training, Ruiz-Olivares et al., 2010; the Promoting Alternative Thinking Strategies curriculum, Domitrovich et al., 2010). These variations should be borne in mind by the reader when interpreting the evidence base discussed below.

## Background evidence

In his 2012/13 annual report, the Chief Inspector for the schools inspectorate raised concerns regarding low-level disruptive behaviour in schools in England (Office for Standards in Education, 2013). A further report suggested that these concerns were shared among parents, carers and teachers (Office for Standards in Education, 2014). For example, two-fifths of c.1000 teachers surveyed by OFSTED identified talking, calling out, and fidgeting as key problems in most lessons, and over a quarter reported that the impact of this low-level disruption on learning was high. These kinds of behaviours are seen as distinct from the more deeply entrenched aggressive, defiant and anti-social behaviours that characterise conduct disorders, which are estimated to affect up to 7% of boys and 3% of girls in childhood and 8% of boys and 5% of girls in adolescence (National Institute for Health and Care Excellence, 2013). However, all behaviour problems are likely to impact upon the learning, participation

and academic achievement of children and young people. In the short term, up to an hour of learning is estimated to be lost each day as a direct consequence of low-level disruption in classrooms (Office for Standards in Education, 2014). In the longer-term, evidence suggests that nascent behaviour problems erode later academic achievement (Gutman & Vorhaus, 2012); this is particularly the case among boys (Panayiotou & Humphrey, 2017).

At the government level, successive Secretaries of State for Education (and those in associated positions) have highlighted disruptive behaviour as a policy priority (for example, while in office, Michael Gove pledged that he would, “not rest when the learning of thousands of children who are desperate to do well and get on is disrupted in classrooms where discipline has broken down”, Haydn, 2014, p.34). Behaviour has been foregrounded in recent iterations of the OFSTED framework (e.g. the ‘Personal development and wellbeing’ strand became ‘Behaviour and safety’, and most recently ‘Personal development, behaviour and welfare’), alongside the production of multiple governmental guidance documents (e.g. Department for Education, 2012, 2014, 2016), and developing the evidence base regarding the most effective behaviour management strategies has been set as a research priority by the Department for Education (2014b). However, the government’s own review of the available data suggests that pupil behaviour is judged to be good or outstanding in the overwhelming majority of schools (e.g. 94% in the primary sector, in which the current study is situated), and that evidence regarding change over time is mixed (Department for Education, 2012b). Thus, it is important to remain sceptical of a ‘crisis model’ of discipline in schools (e.g. *Behaviour is a national problem in schools in England*, The Guardian; Grierson, 2017).

## **Impact of the GBG**

### *Behavioural outcomes*

There exists a plethora of studies attesting to the positive effects of the GBG on children’s behaviour (Donaldson & Wiskow, 2017; Embry, 2002). More specifically, researchers have examined its impact on *challenging* behaviours, including for example aggressive (e.g. hitting others), disruptive (e.g. talking out), and off-task (e.g. failing to pay attention) behaviours (Flower et al., 2014). Additionally, although studied less frequently, the ability of the GBG to increase appropriate, pro-social and on-task behaviours has also been documented (Tingstrom et al., 2006). Flower et al.’s (2014) meta-analysis of 22 GBG studies provides a useful summary of the evidence in this regard; these authors found a moderate ( $d = .50$ ) average effect on behavioural outcomes. To place this in context, three systematic reviews and meta-analyses (Korpershoek, Harms, de Boer, van Kuijk, & Doolaard, 2016; Oliver, Wehby, & Reschly, 2011; Whear et al., 2013) covering a wide range of classroom management strategies and practices found small average effects on behavioural outcomes (e.g.  $g = 0.24$  in the Korpershoek et al. analysis). Of particular note is that the Korpershoek et al (2016) analysis proffered a much more conservative estimate of the GBG’s effects ( $g = 0.25$ ), though this may be due to their more robust criteria<sup>3</sup>, which led to the inclusion of only four studies focusing specifically on this intervention.

### *Academic outcomes*

Despite the proliferation of the GBG, there remains a distinct lack of robust and consistent evidence available regarding its impact on academic attainment (Weis et al., 2015). Bradshaw, Zmuda, Kellam, and Ialongo’s (2009) longitudinal study reported positive effects on a range of academic outcomes (including standardised achievement tests) at age 19 following a single year of intervention exposure at age 6-7. The results of this RCT are promising, with standardised effect sizes equating to four and five months of additional progress in reading and maths respectively. However, it is important to note

---

<sup>3</sup> The Korpershoek et al. (2016) meta-analysis included only matched quasi-experimental and RCT designs, whereas the Flower et al. (2014) meta-analysis also included single subject experimental designs.

that the children in the intervention arm of this study were also in receipt of an intensive “enhanced academic curriculum” (ibid, p.6), making it very difficult to ascribe these effects directly and solely to the GBG. Other trials have addressed this issue. However, the first of these, reported by Dolan et al (1993) found that the exposure to the GBG for two years at age 6-8 did not impact upon reading outcomes in the short-term; Kellam et al. (2008) and Hemelt, Roth, and Eaton’s (2013) long-term follow up studies of this sample also found null results at the intention-to-treat (ITT) level in relation to high school and college outcomes respectively. The second trial, by Dion et al. (2011), examined the effects of a single year of exposure to the GBG combined with peer tutoring (compared to peer tutoring only and usual practice) on a range of literacy outcomes (e.g. reading comprehension) among 6-7 year olds. The authors found significant improvements in literacy following exposure to peer tutoring, but not the GBG. It is also worthy of note that the authors of this study applied selection criteria in the recruitment stage of the trial, leading to, “an overrepresentation of inattentive students at-risk of reading problems” (p.72), thus limiting the generalisability of findings to the broader school population. Finally, a recent study by Weis et al (2015) reported small effects of the GBG on children’s reading and mathematics scores after one year of implementation; however, the quasi-experimental (as opposed to randomised) design of this study limits the security of its findings.

In sum, there is genuine uncertainty regarding the potential of the GBG to improve children’s academic outcomes. Thus, a key intended contribution of the trial reported here is to provide robust evidence in this regard.

#### *Differential gains*

Although ITT analyses in an RCT provide the most unbiased estimate of the impact of an intervention (Gupta, 2011), it is known that participants do not respond in a uniform manner (Farrell, Henry, & Bettencourt, 2013), and as such the ITT approach may underestimate impact by failing to appreciate the natural diversity in universal populations (Greenberg & Abenavoli, 2017). However, there is also a valid concern that subgroup moderator analyses can introduce bias and lead to over-interpretation of intervention effects (Petticrew et al., 2012). Thus, such analyses can be useful as a supplement to ITT provided that they are specified in advance and consistent with theory and evidence pertaining to the intervention (Humphrey et al., 2016). Differential gains following exposure to the GBG among population subgroups have been examined primarily in relation to gender (e.g. Vuijk, van Lier, Crijnen, & Huizink, 2007), baseline levels of challenging behaviours (such as elevated aggression, Dolan et al., 1993), and the interaction between the two (e.g. amplified intervention effects among aggressive boys, Kellam, Rebok, Ialongo, & Mayer, 1994). These effects are theoretically plausible given the gendered socialisation of competitiveness (Gneezy, Leonard, & List, 2009) and responses to social task demands in the classroom (Kellam et al., 1994); as such, we sought to undertake a *confirmatory* subgroup analysis (Varadhan, Segal, Boyd, Wu, & Weiss, 2013) relating to boys with elevated levels of challenging behaviours at baseline.

In relation to the primary remit of the EEF – improving outcomes among children and young people from socio-economically disadvantaged backgrounds - there has been only very limited exploration of potential differential gains following exposure to the GBG, and that research which has been conducted provides equivocal evidence. Thus, while Spilt, Koot, and Lier (2013) included low socio-economic status (SES) in their wide-ranging and comprehensive analysis of subgroup differences in the impact of the GBG on internalizing and externalising behaviours, their analytical approach (a person-centred approach in which numerous risk variables at different ecological levels contributed to six different risk profiles) precluded precise estimation of subgroup effects *specifically* among low SES students. Furthermore, while the aforementioned study by Weis et al (2015) found amplified effects of the GBG among students who attended low and moderate (as opposed to high) SES schools, their analysis did not account for SES at the individual level. Finally, while Kellam, Ling, Merisca, Brown, and Ialongo (1998) included both classroom and individual level SES in their models examining the effects of the GBG, they did not explicitly examine whether said effects varied as a function of either of these

variables. However, from a theoretical point of view, it is plausible that school-based interventions may compensate for some of the factors that constrain the academic achievement of students from socio-economically disadvantaged backgrounds (Dietrichson, Bøg, Filges, & Klint Jørgensen, 2017), and indeed there is some tentative evidence of differential gains among this subgroup following participation in other universal preventive interventions (e.g. Second Step, Holsen, Iversen, & Smith, 2009). Given this, we sought to undertake an *exploratory* subgroup analysis (Varadhan et al., 2013) relating to students eligible for free school meals (FSM).

### *Teacher outcomes*

The primary focus of research on the impact of the GBG has been on pupil outcomes. However, there exists the potential for this intervention to also create meaningful impact on certain teacher outcomes (Elswick & Casey, 2011). In framing this aspect of the study, we draw on Jennings and Greenberg's (2009) 'pro-social classroom' model, which posits that teacher wellbeing, teacher-pupil relationships, classroom management, effective implementation of preventive interventions, and pupils' social, emotional, behavioural and academic outcomes, are reciprocally inter-related. Assuming that the GBG is implemented well and positively impacts upon pupils' behaviour, even if not their academic attainment (see above), it is plausible that the experience of implementation and subsequent observed changes in behaviour will increase teachers' sense of self-efficacy in classroom management (Kelm & McIntosh, 2012). Furthermore, given the established association between pupil behaviour and teacher stress (McCormick & Barnett, 2011), one could also predict reductions in teacher stress (although this hinges on the assumption that the intervention produces observable changes in behaviour that are meaningful to participating teachers). However, it should be noted that one could also predict the converse – that is, the GBG increases task load for teachers and thus may actually lead to higher levels of stress. Finally, research which highlights the connection between pupil behaviour stressors and teacher attrition (e.g. Sass, Seal, & Martin, 2011) indicates that effects on retention outcomes are also an avenue worthy of exploration. However, expectations relating to this outcome are necessarily tempered by the knowledge that factors such as workload and policy changes are the most powerful drivers of teacher attrition (Department for Education, 2017).

The above predictions remain almost completely untested to date. Existing research on GBG teacher outcomes has been restricted to those which might be considered somewhat 'treatment-inherent' (Slavin & Madden, 2011) in that they focus on behavioural changes connected directly to the nature and content of the intervention (e.g. increased use of verbal praise to reinforce appropriate behaviours, Elswick & Casey, 2011; Lannie & McCurdy, 2007; Lynch & Keenan, 2016). To date only Domitrovich et al (2016) have rigorously examined the impact of the GBG on treatment-independent outcomes such as those noted above. In their three arm RCT, these authors found that the GBG combined with the PATHS curriculum, but not the GBG alone, led to significant increases in teachers' self-efficacy in social and emotional learning, behaviour management, and personal accomplishment, when compared to a control group of teachers not implementing either intervention. This has led us to undertake an exploratory analysis of the impact of the GBG on teachers' self-efficacy in classroom management, stress, and retention.

### **Implementation and process evaluations of the GBG**

As is the case for many evaluations of school-based interventions, the overwhelming majority of data collected pertaining to implementation of the GBG has been used descriptively as a means to increase internal validity of trial findings (see for example Dion et al., 2011; Ialongo, Poduska, Werthamer, & Kellam, 2001). To date, there has been virtually no examination of the extent to which different levels of GBG implementation (e.g. within dimensions such as fidelity, quality, and dosage) are associated with outcome variability. One exception is Ialongo et al. (1999), whose analyses demonstrated that higher fidelity to intervention protocols was associated with greater impact on behavioural and academic outcomes. The current study provides an opportunity to build upon and extend this work by assessing

the relative influence of a range of implementation dimensions (including, but not limited to fidelity) on intervention outcomes. This was considered to be particularly pertinent given the strong emphasis on specific aspects of implementation in the GBG (e.g. procedural fidelity and dosage – see ‘Intervention’ subsection above), and in light of the broader evidence base in which the moderating influence of implementation variability on the outcomes of preventive interventions is well established (Durlak, 2016).

In spite of the lack of implementation-outcomes analyses, existing implementation and process (IPE) research pertaining to the GBG can still provide useful information in the context of the current study. Of particular note, reported levels of fidelity are generally high but do vary across studies (e.g. 82% in Domitrovich et al., 2015; 92% in Dion et al., 2011; 60% in Jalongo et al., 2001; 77% in Leflot et al., 2013). Reporting of dosage data is surprisingly limited given the nature of the GBG (e.g. the emphasis given to the frequency and duration of games). Hagermoser-Sanetti and Fallon (2011) reported that although 94% of teachers played the GBG for the recommended duration, only 56% played it at the recommended frequency, and only 31% implemented the recommended number of probe sessions. This contrasts with the findings of Domitrovich et al (2015), who found that the average frequency with which games were played (daily) was in line with expectations, the average duration of games (<10 minutes) was not. Hence, where data are available, they suggest that GBG implementation is variable across studies and often not in line with developer expectations. This is, of course, a common and longstanding finding in implementation research (Lendrum & Humphrey, 2012). While not particularly surprising then, these data do prompt questions about the drivers of this implementation variability. Perceptions of social validity (e.g. acceptability, feasibility, utility) appear to be generally positive in relation to the GBG (Kleinman & Saigh, 2011; Tingstrom, 1994), and there is little evidence that any teacher (e.g. demographic or professional) or organisational (e.g. school structural and compositional) characteristics yield any influence on implementation (Domitrovich et al., 2015). Thus, we sought to provide further clarification on factors affecting the implementation of the GBG as part of the IPE strand of the current study.

### **The UK evidence base**

The UK evidence base for the GBG is currently extremely sparse, with only three published studies. The first and second of these examined the utility of the intervention in reducing off-task behaviour of children and adolescents attending special schools (Phillips & Christie, 1986; Webster, 1989). However, the age of these studies, the extremely small scale of the research reported, the specialist setting and focus on students with special educational needs (SEND), and the lack of a comparison group in either study precludes any firm conclusions from being drawn regarding the likely impact of the GBG in UK mainstream schools.

More recently, the intervention was piloted in six Oxfordshire primary schools (10 classes, N=222 children, aged 5-9) over the course of one year (Chan et al., 2012; Coombes et al., 2016). The evaluation of this pilot indicated that implementation of the GBG was associated with significant improvements in a range of behaviours (e.g. attention/concentration) assessed by the Teacher Observation of Classroom Adaptations (Revised) scale (TOCA-R) (Werthamer-Larsson, Kellam, & Wheeler, 1991). Furthermore, analysis of game and probe data indicated a clear trend indicative of generalisation (e.g. reductions in rule infractions during probe sessions) across the course of the year. As in the earlier UK studies though, the Oxfordshire pilot did not include a control group, thus limiting the extent to which these behavioural improvements could be securely attributed to the GBG.

However, the Oxfordshire pilot did highlight several implementation and process-related issues that are salient in the context of the current study. In particular, challenges were identified in relation to implementation burden (e.g. required workload, time involved in preparing for and playing the game), integrating the GBG with other teaching, and aspects of the intervention that were seen to clash with established and preferred teaching styles (e.g. the requirement to not directly interact with individual



pupils during game sessions). In spite of these challenges, teachers reported a range of benefits for pupils in their classes, but also highlighted the importance of coaches in supporting high quality implementation. Ultimately, the authors concluded that the GBG was both feasible and acceptable to English primary school teachers and head-teachers (Chan et al., 2012; Coombes et al., 2016).

## Evaluation objectives

Our team conducted a major efficacy trial of the GBG in England that focused on (i) the intervention's impact on children's educational outcomes (e.g. reading, behaviour); and in particular (ii) its impact on boys displaying borderline/abnormal levels of conduct problems; and (iii) children eligible for FSM; (iv) examining whether the way in which the GBG is implemented is associated with variability in outcomes; and (v) whether the GBG improves outcomes for teachers (specifically, self-efficacy in classroom management, classroom stress, and retention). The study protocol can be found [here](#).

## Hypotheses

H1: Children in primary schools implementing the GBG over a two-year period will demonstrate significant improvements in reading (1a) and behaviour (specifically, concentration problems - 1b; disruptive behaviour - 1c; and, pro-social behaviour - 1d) when compared to those children attending control schools.

H2: The effects outlined in H1 above will be amplified for boys exhibiting borderline/abnormal levels of conduct problems at baseline.

H3: The effects outlined in H1 above will be amplified for children eligible for free school meals.

H4: Variation in implementation fidelity/quality (4a)<sup>4</sup>, dosage (4b), reach (4c), and participant responsiveness (4d), will be significantly associated with reading and behavioural outcomes among pupils in schools implementing the GBG.

H5: Teachers implementing the GBG will demonstrate measurable improvements in self-efficacy in classroom management (5a), classroom stress (5b), and retention (5c), when compared to teachers in control schools.

## Ethical review

The study was approved by the University Research Ethics Committee at the University of Manchester (Ref: 15126).

Consent/assent involved three stages. Firstly, participating schools signed a Memorandum of Agreement (MoA) indicating their willingness to participate. The MoA contained detailed information about what participation entailed (e.g. data collection procedures and requirements, plus payment of a contributory fee for those schools randomly allocated to the intervention arm). In addition, it explained the nature of the RCT (e.g. that only half of participating schools would receive the GBG, and that this would be determined by a random allocation procedure), and what schools could expect in return for their participation (e.g. aggregated survey feedback, plus a nominal fee for compliance with data collection requirements among schools randomly allocated to the usual provision arm). Secondly, participating schools distributed consent forms to the parents and carers of all eligible pupils, specifically, pupils in Year 3 classes during the academic year 2015/16. Parents and carers who did not want their child to participate in the trial completed an 'opt-out' section on the consent form which was

---

<sup>4</sup> Fidelity and quality have been conflated in this hypothesis in light of the results of a factor analysis of our observational data – see 'structured observations' subsection in Methods chapter.

returned via a freepost address to the University of Manchester. In total, 68 parents (2.2%) exercised their right to opt their children out of the trial. Finally, children were provided with information about the study (including their guarantee of anonymity and right to withdraw) and were asked to give their assent to participate. No children declined assent or exercised their right to withdraw from the study.

An additional consent process was followed for any children attending case study schools in the IPE strand of the trial who were nominated to participate in the pupil focus groups. This followed an explicit (e.g. opt-in, as opposed to opt-out) parental consent procedure. Standard protocols were followed in respect to confidentiality and disclosure during the conduct of these focus groups. Children were assured that their responses would remain anonymous and confidential except in the event of the disclosure of information indicative of a child protection issue, at which point the school's designated safeguarding lead would have to be informed. No such disclosures took place.

Anonymity and confidentiality was ensured through data management procedures as follows: security for online surveys was ensured using hypertext-transfer-protocol-secure data encryption. Data matching (e.g. across time) was achieved through the use of a unique pupil number. All qualitative data were anonymised during the transcription process, with pseudonyms given to any personally identifying information. The University of Manchester and Microsoft Best Practice guidelines for data storage were followed, ensuring that data was held safely on secure drives behind internal and external firewalls, and physical transportation prohibited (e.g. flash drives).

### **Project team**

#### **Delivery team**

Kate O'Brien: Director of Programmes - responsibility for GBG delivery

Michael O'Toole: Chief Executive - responsibility for GBG delivery, oversaw relationship with EEF

Simon Claridge - Director of Programmes during recruitment phase.

Alessandra Podesta: GBG programme manager

Amanda Hood: Administrator

Peter Wilde: Head GBG coach

Lauren Bond: Head GBG coach

Emma Dove: GBG school coach

Steve Iredale: GBG school coach

Kate Gummatt: GBG school coach

Carol Healy: GBG school coach

John Killeen: GBG school coach

Kirsty Pert: GBG school coach

John Rees: GBG school coach

#### **Evaluation team**

Neil Humphrey: principal investigator and lead author

Alexandra Hennessey: trial manager and quantitative analyst

Emma Ashworth: research assistant and qualitative analyst

Kirsty Frearson: research assistant and qualitative analyst

Louise Black: research assistant

Kim Petersen: research assistant

Lawrence Wo: data manager

Margarita Panayiotou: quantitative analyst

Ann Lendrum: specialist in implementation and process evaluation

Michael Wigelsworth: specialist in assessment of outcomes

Liz Birchinnall: specialist in primary education

Garry Squires: specialist in mental health and therapeutic intervention in schools

Maria Pampaka: specialist in measurement and survey design

### **Trial registration**

The trial was registered with ISRCTN (Ref: 64152096, details [here](#)).

## Methods

### Trial design

A two-year cluster-randomised trial design was used, with schools as the unit of randomisation. This design is advantageous in terms of the balance between scientific rigour, ethical considerations, and goodness-of-fit with the study aims and hypotheses. Schools were the unit of randomisation in order to minimise the risk of contamination that would have been associated with within-school (e.g. class) randomisation, and also to reflect the practical consideration that the intervention model includes a GBG coach being assigned to each participating school in the intervention arm.

Schools were randomly allocated to one of two trial arms: (1) to deliver the GBG for a subsidised fee of £1,500<sup>5</sup> (intervention arm); or (2) continue as normal and receive financial compensation of £1,500 for participating in data collection (usual provision arm). Teachers in schools allocated to the intervention arm were trained and supported to implement the GBG during the two-year trial period (2015/16 and 2016/17). Their counterparts in schools allocated to the usual provision arm continued their normal practice during the same period. At the conclusion of the trial, schools were free to decide whether to continue (in the case of the intervention schools) or to start (in the case of usual provision schools) implementing the GBG.

### Participant selection

Eligibility to participate required schools to be state-maintained and not already implementing the GBG. Three regions (Greater Manchester, West and South Yorkshire, and the East Midlands) were targeted for recruitment. The school recruitment process was primarily handled by the grantee (Mentor UK) project team, who employed a number of strategies, including holding regional recruitment events, using contacts at Local Authorities and independent providers (e.g. One Education) to identify prospective trial schools, and emailing project flyers to schools. Initial expressions of interest were sought via an online form, followed by direct contact from Mentor UK and/or the research team, leading ultimately to the signing of the aforementioned MoA. Participating schools signed the MoA, paid an average of £1,500<sup>5</sup>, distributed information and consent sheets to parents of pupils in the target cohort, and completed a minimum of 90% of baseline surveys prior to randomisation. Schools allocated to the control group had their £1,500<sup>5</sup> fee returned, and received a further £1,500<sup>5</sup> in two instalments (£1,000 initially, £500 at the end of the trial) if they complied with data collection requirements.

In total, 77 primary schools complied with the above requirements, of which 38 were randomly allocated to the GBG arm and 39 to the usual provision arm. The target cohort were pupils in Year 3 classes in the first year of the trial (2015/16). After accounting for parental opt-outs (n=68, 2.2%), this cohort consisted of N=3,084 pupils. Copies of the Expression of Interest form, Memorandum of Agreement, and parental information sheet and consent form can be found in Appendices 1-3.

At T1 (summer term 2015) only, we also used the teacher-rated conduct problems subscale of the Strengths and Difficulties Questionnaire (SDQ; Goodman, 1997) in order to identify the at-risk sample of boys for H2. This 5-item scale requires raters to read statements about a child's behaviour (e.g. "Often has temper tantrums or hot tempers") and endorse them on a 3-point scale (not true/somewhat true/certainly true). The subscale produces a score of 0-10, with 0-2, 3 and 4-10 representing the normal, borderline and abnormal ranges respectively. At-risk status was defined as scoring in the borderline or abnormal range on this measure at T1. The conduct problems subscale of the SDQ has

---

<sup>5</sup> Variable by school size.

satisfactory internal consistency ( $\alpha=0.70$ ), test–retest reliability ( $r=0.77$ ) and the capacity to discriminate between children with and without psychiatric diagnoses (Stone, Otten, Engels, Vermulst, & Janssens, 2010). Internal consistency of this subscale in the trial was very good ( $\alpha=0.80$  at baseline).

## Outcome measures

Outcomes were assessed annually at baseline (T1, summer term 2015), the end of the first year of the trial (T2, summer term 2016), and at the conclusion of the trial (T3, summer term 2017). The primary outcome measure for this study was children’s attainment in reading. Secondary outcome measures at the pupil-level were children’s disruptive behaviour, concentration problems and pro-social behaviour. At the teacher-level, we assessed self-efficacy in classroom management, classroom stress, and retention. All outcome measures are described in more detail below.

The primary outcome was assessed at T1 and T3 only. Pupil-level secondary outcomes were assessed at T1, T2 and T3 (T2 being used solely for interim analyses)<sup>6</sup>. Teacher-level secondary outcomes were assessed at T1 and T2 for those teaching Year 3 classes in 2015/16, and at T2 and T3 for those teaching Year 4 classes in 2016/17<sup>7</sup>. With the exception of the primary outcome measure, all of the above were administered online at a secure, password-protected site powered by the World App Key Survey platform.

### Primary outcome

The primary outcome measure at baseline (T1) used data from the National Pupil Database end of Key Stage 1 teacher assessments (specifically KS1 National Curriculum reading point score: the KS1\_READPOINTS variable). This was collected as part of national tests across England in spring 2015. For the follow-up (T3) assessment, we used the Hodder Group Reading Test (HGRT; specifically, test sheet 2A). This paper-based measure was developed for use from age 7-12 years, takes a maximum of 30 minutes to complete (thus minimising data burden), and was standardised on over 13,000 pupils. It is administered in a whole-class/group context, utilises a multiple-choice response format, and assesses children’s reading comprehension at word, sentence and continuous text level ([www.hoddertests.co.uk](http://www.hoddertests.co.uk)). The HGRT produces raw scores that can be transformed into National Curriculum levels, reading ages and standardised scores (NB: our primary outcome analysis used raw scores). Members of the research team administered the test at T3 (summer term 2017). We could not employ a fully blinded approach as the individuals involved were already aware of the allocation status of individual schools as a result of their previous contact with them at earlier stages in the trial (and, of course, the various physical artefacts associated with the GBG such as the rules posters would have spoiled any blinding). This was not considered a serious risk as the evaluation team was independent and not invested in the intervention. Every test paper was double-marked by members of the research team to eliminate human error. In instances where discrepancies were found, those were eliminated via joint reference to the scoring protocol.

### Secondary outcomes

#### *Disruptive behaviour, concentration problems and pro-social behaviour*

Children’s behaviour was assessed using the 21-item Teacher Observation of Children’s Adaptation checklist (TOCA-C; Koth, Bradshaw, & Leaf, 2009). Teachers read statements about a child (e.g. “Pays

---

<sup>6</sup> In addition, teachers completed the conduct problems subscale of the Strengths and Difficulties Questionnaire (Goodman, 1997) at T1 to facilitate the subgroup analysis for H2.

<sup>7</sup> Completion of assessments of teachers’ usual practice for the IPE strand of the trial also followed this pattern.

attention”) and endorse them on a 6-point scale (never/rarely/sometimes/often/very often/almost always). The disruptive behaviour subscale includes items reflecting disobedient, disruptive and aggressive behaviours. The concentration problems subscale includes items reflecting inattentive and off-task behaviour. Finally, the pro-social behaviour subscale includes items reflecting positive social interactions. The TOCA-C is internally consistent (all subscales  $\alpha > 0.86$ ) and has a factor structure that is invariant across gender, race and age (Koth et al., 2009). Internal consistency of the TOCA-C subscales in the trial was excellent (all  $\alpha > 0.87$  at baseline).

#### *Teacher efficacy in classroom management*

Teacher efficacy in classroom management was assessed using the 4-item subscale of the short-form Ohio State Teachers’ Sense of Efficacy Scale (OSTES; Tschannen-Moran & Hoy, 2001). Teachers read questions (e.g. “How much can you control disruptive behaviour in the classroom?”) and respond on a 9-point scale with five equally spaced anchors (not at all/very little/some influence/quite a bit/a great deal). The classroom management subscale of the OSTES is internally consistent ( $\alpha = 0.86$ ) and its criterion validity has been established (e.g. correlates significantly with established teacher efficacy measures) (Tschannen-Moran & Hoy, 2001). Internal consistency of this subscale in the trial was excellent ( $\alpha = 0.90$  at T1, summer term 2015).

#### *Teacher classroom stress*

Teacher stress was captured using the 5-item pupil misbehaviour subscale of the Teacher Stress Inventory (TSI; Boyle, Borg, Falzon, & Baglioni, 1995). Respondents read questions (e.g. “How great a source of stress is maintaining class discipline?”) and respond on a 5-point scale (no stress/mild stress/moderate stress/much stress/extreme stress). The pupil misbehaviour subscale of the TSI is internally consistent ( $\alpha = 0.77$ ) and its construct validity has been demonstrated (e.g. scores for secondary school teachers are significantly higher than for primary school teachers) (Borg, Riding & Falzon, 1991). Internal consistency of this subscale in the trial was very good ( $\alpha = 0.82$  at T1, summer term 2016).

#### *Teacher retention*

Consistent with recent research on teacher workload (e.g. Lightfoot, 2016), retention was assessed through the use of a single item measure, as follows: “How likely are you to leave the teaching profession in the next 5 years?” Participating teachers responded on a 6-point scale (definitely/highly likely/likely/unlikely/highly unlikely/definitely not).

## **Sample size**

### **Calculation of sample size**

Sample size calculations were carried out using Optimal Design Software. Initial calculations, published in the study protocol, identified a need for a *minimum* of 72 schools and an estimated sample of 2,880 pupils (an average of 40 pupils per school – based on the proportion of single/mixed, double and triple form entry schools recruited in our previous EEF trial; Humphrey et al., 2015). Using a demographic and pre-test covariate model, we assumed an intra-cluster correlation co-efficient (ICC) of no more than 0.06 for our primary outcome measure (Hedges & Hedberg, 2007). Given this, and standard Power and Alpha thresholds of 0.80 and 0.05 respectively, the trial would be powered for a minimum detectable effect size (MDES) of 0.20 in an ITT analysis for H1. This sample size would also be powered for MDESs of 0.37 in the at-risk boys subsample (H2; initial assumption of  $n = 258$ , 9%) and 0.25 in the FSM subsample (H3; initial assumption of  $n = 864$ , 30%).

As noted above, 77 schools (N=3,084 pupils) ultimately participated<sup>8</sup>. The number of at-risk boys (H2) was determined to be n=337 (11% of the trial sample). The number of children eligible for FSM (H3) was determined to be n=764 (24.8%). In the interests of clarity, the MDES at protocol, randomisation and analysis stages is presented in Table 3 (see Impact Evaluation chapter).

## Randomisation

The random allocation procedure was conducted independently of the research team by the Manchester Academic Health Science Centre Clinical Trials Unit. A minimisation algorithm was utilised to ensure balance across the arms of the trial in terms of the proportion of children eligible for FSM and school size<sup>9</sup>, using data from the school performance tables on the DfE website.

## Analysis

ITT analyses were conducted for the primary and secondary pupil-level outcomes using raw data in all cases. Multi-level models (MLM) with fixed effects and random intercepts in MLwiN2.36 were used. Two-level (school, pupil) hierarchical models, controlling for baseline (T1) scores at the pupil level, were fitted to account for the nested nature of the data. Follow-up (T3) outcome scores were used as the response variable. Initially, empty ('unconditional') models were fitted, entering only the school identifiers and no explanatory variables, in order to allow approximations of the proportion of unexplained variance attributable to each level of the model. A full ('conditional') model was then fitted, entering trial group (GBG vs. usual provision) at the school level, and baseline (T1) score at the pupil level. An intervention effect was noted if the co-efficient associated with the trial group variable was statistically significant. This was subsequently converted to Hedge's *g* accounting for varying cluster sizes, as per EEF reporting guidelines (Hedges, 2007; Tymms, 2004). The coefficients reported are based on raw scores. The standardised effect size, Hedges *g*, was calculated using the coefficient of the trial group effect divided by the square root of the pooled pupil-level variance (the square root of the within group variance) from an empty model (Tymms, 2004). The pupil-level variance was used as the hypothesized impact of the intervention is on pupil level outcomes. Confidence intervals were calculated as the effect size +/- the product of the critical value of the normal distribution ( $\approx 1.96$ ) and the standard error of the group indicator coefficient (standardised) estimated from the MLMs.

Fulfilment of our study objectives necessitated planned subgroup analyses (e.g. H2, H3). For H2, the MLMs outlined above were extended to include gender and risk status at the pupil-level, and the cross-level interaction term 'group\*risk\*gender' (if GBG, if at-risk, if male). For H3, the same procedure was applied, with models extended to include FSM<sup>10</sup> at the pupil-level, and the cross-level interaction term 'group\*FSM' FSM (if GBG, if FSM eligible). An intervention effect at the subgroup level was noted if the coefficients associated with these interaction terms were statistically significant. Conversion of raw score coefficients to Hedge's *g* followed the same procedure as noted above.

Our principal analyses used fully observed data, as per EEF guidelines (Model 1.1). Subsequently, we assessed the sensitivity of our findings to the inclusion of minimisation variables at the school level, and FSM eligibility and gender at the child level (Model 1.2). Models were then extended to include the above noted interaction terms pertaining to at-risk boys (Model 1.3) and children eligible for FSM (Model 1.4), ahead of a final model in which all explanatory variables used in the preceding analyses were

---

<sup>8</sup> This over-recruitment was undertaken in order to allow for expected school-level attrition through the course of the trial.

<sup>9</sup> Schools were split into terciles for low, moderate and high proportions of FSM and size.

<sup>10</sup> NB: Current FSM eligibility was used as opposed to 'everFSM' as the NPD request for the trial preceded the EEF decision to adopt everFSM as its preferred FSM variable.

included simultaneously (Model 1.5). The sensitivity of these findings to the use of both fully *and* partially observed data via multiple imputation (MI) was then assessed (Models 2.1 to 2.5, respectively). For further details on missing data and MI please refer to the 'missing data' subsection of the Impact Evaluation chapter.

For teacher-level outcomes (H5), single level linear regression models were fitted in SPSS version 22 as follows: follow-up score (T2 for Year 3 teachers, T3 for Year 4 teachers) as the response variable, with baseline score (T1 for Year 3 teachers, T2 for Year 4 teachers) and trial arm (GBG vs. usual provision) as explanatory variables. To account for missing data, subsequent models using Mplus8 were tested using maximum likelihood (Full Information Maximum Likelihood) with robust standard errors (Robust Maximum Likelihood). A copy of the approved statistical analysis plan for the trial can be found [here](#).

## Implementation and process evaluation

The IPE strand of the study comprised three components. Firstly, all teachers of children in the trial were surveyed about their classroom behaviour management practices in order to establish a clear counterfactual (e.g. what does 'usual provision' in the control group look like?) and give an indication of the level of programme differentiation (e.g. to what extent is the GBG distinct from existing behaviour management practices?). Secondly, independent structured observations of teachers' implementation of the GBG were conducted, focusing on fidelity (e.g. to what extent do teachers adhere to prescribed procedures when playing the game?), quality (e.g. how well do teachers deliver the components of the GBG?), participant responsiveness (e.g. to what extent do children engage with the GBG?), and reach (e.g. what is the rate and scope of participation in the GBG across the class?). The data generated through these observations were used alongside the aforementioned dosage data derived from the online scoreboard to provide summative descriptions of GBG implementation through the course of the trial and also to assess the extent to which different levels of these implementation dimensions were associated with outcome variability (H4). Thirdly, we conducted longitudinal case studies of six GBG schools as a means to provide a rich, detailed picture of the implementation process and the factors underpinning it, using a social validity framework (e.g. acceptability, feasibility, utility). Case study data were generated via interviews, focus groups, observations and document analysis, drawing upon the views of a range of informants (e.g., pupils, teachers, school leaders, parents, GBG coaches).

Each of the above components is discussed in more detail below. First, however, we draw the reader's attention to the discontinuation of implementation among several GBG schools during the trial,, in order that the information that follows regarding other aspects of the IPE methodology make sense (e.g. the number of structured observations conducted). In total nine of the 38 GBG schools (24%) ceased implementation before the conclusion of the trial<sup>11</sup>, including two of the longitudinal case study schools noted above. Three schools stopped implementing the intervention in 2015/16 (n=4 classes, plus n=1 class that discontinued in a school where other classes continued), and a further six ceased in 2016/17 (n=12 classes), plus n=1 class that discontinued in a school where the other class continued. Two of these schools agreed to take part in 'exit interviews', which were supplemented by analysis of email communication between the other schools and the project delivery team (Mentor UK) in order to develop our understanding of the factors that contributed to their decision. Our analyses of these data are reported in the Implementation and Process Evaluation chapter.

---

<sup>11</sup> NB: All schools continued to comply with trial outcome data collection requirements; hence, there was no school-level attrition in the trial (see CONSORT diagram in Impact Evaluation chapter).



### Usual practice survey

Teacher usual practice surveys were administered alongside teacher-level outcome surveys following the pattern noted earlier (see 'outcome measures' above). Based on an existing measure of teachers' classroom management strategies (Reupert & Woodcock, 2010), the survey consisted of 52 items assessing general demographic information (3 items) and practice in three domains: general behaviour management approaches (22 items, e.g. "I establish and maintain a set of classroom rules", Yes/No response format), use of reward systems (10 items, e.g. "I use group rewards", Never/Monthly/Weekly/Every Day) and approaches to managing disruptive and inappropriate behaviour (17 items, e.g. "I use a warning/strike system", Never/Monthly/Weekly/Every Day). The structure of each domain was explored using Mplus 8. A parallel analysis with 5000 random datasets was conducted to assess how many factors to retain<sup>12</sup>. Exploratory Factor Analysis (EFA) with Weighted Least Squares Mean and Variance adjusted (WLSMV) was then used to assess their structure, while accounting for the clustering in the data. Only items with factor loadings above .32 were retained (Tabachnick & Fidell, 2013). Due to sample size limitations, a Confirmatory Factor Analysis (CFA), commonly used to establish the structure found in EFA, was not feasible. Analyses indicated a one-factor structure for general behaviour management (20 items), one-factor structure for reward systems (8 items,  $\omega=0.72$ ), and a two-factor structure for managing disruptive behaviour (these were tentatively named 'use of physical and verbal reprimands', and 'systems and procedures for managing disruption', comprising 5 items,  $\omega=0.61$ , and 9 items,  $\omega=0.72$ , respectively) offered the best fit to the data. See Appendix 4 for more details.

### Structured observations

Observational data were utilised for H4 as these are widely considered to be the most rigorous source of implementation data (Humphrey et al., 2016); by contrast, implementer self-report data can be positively biased, being subject to demand characteristics and impression management. The development of the structured observational schedule was informed by those used in previous GBG studies (e.g. Leflot, van Lier, Onghena, & Colpin, 2013), the GBG implementation manual and fidelity checklist published by AIR (Ford, Keegan, Poduska, Kellam, & Littman, 2014), our own work in other trials (e.g. PATHS trial, Humphrey et al., 2015), and the extant literature on implementation and process evaluation (e.g. Hansen, 2014). A draft of the schedule and accompanying explanatory rubric was developed by the evaluation team ahead of piloting and refinement using video footage of the GBG being implemented in English schools in the UK pilot (Chan et al., 2012; Coombes et al., 2016). In this initial formative stage, which lasted several days, the emphasis was on aligning our understanding of the various implementation indicators and their application in the context of the GBG. Given the use of multiple observers (N=3), additional video footage of GBG implementation was then used in order to generate inter-rater reliability data for each indicator. These analyses demonstrated exceptionally good inter-rater reliability (e.g. Cohen's Kappa for our nominal procedural fidelity items was 0.95, indicative of near perfect agreement). A copy of the observation schedule can be found in Appendix 5.

Each class in the GBG arm of the trial was observed twice (once in each year, with the exception of classes in schools/classes that had ceased implementation, n=6 in 2015/16, and a further n=12 in 2016/17, as noted above). In 2015/16, 54 of the 60 classes were observed playing the GBG (five classes had ceased implementation, and one observation could not be arranged). In 2016/17, 46 of the 58

---

<sup>12</sup> Parallel analysis is considered to be one of the most robust methods for deciding the number of factors to retain in a measure. It involves extracting eigenvalues from k random datasets that parallel the actual data. The 50<sup>th</sup> and 95<sup>th</sup> percentile of the eigenvalues derived from random data are then compared to the eigenvalues of the actual data. Factors retained are those with eigenvalues greater than that of the random data (50<sup>th</sup> and 95<sup>th</sup>). This ensures that the factors retained can account for more variance than what would be expected by chance alone.

classes were observed playing the GBG (14 classes had ceased implementation). In total 100 classes from 35 GBG schools were observed over the course of the trial. In order to streamline analyses, and thus reduce the likelihood of model overfitting and avoid collinearity, the observer-rated implementation data were subjected to EFA using Mplus8 (see Appendix 6). Given the small sample size, procedural fidelity items were summed to represent pre-game, during game, and post-game fidelity (each scored from 0-100%); these were included in the EFA alongside implementation quality (5 items) and participant responsiveness (5 items) indicators. The same analytic procedures outlined above in relation to the usual practice survey were applied to this dataset. These analyses indicated a two-factor structure for the observational data (fidelity/quality,  $\omega = .66$ , and participant responsiveness,  $\omega = .72$ , respectively). This distinction between the behaviour of the *implementer* and that of the *recipients* of the intervention is consistent with implementation theory (e.g. Berkel, Mauricio, Schoenfelder, and Sandler's (2011) integrated model of implementation). The two factors were supplemented by data on participant reach (derived from a log of the proportion of pupils in a given class that were present while the game was being played) and dosage (derived from the online scoreboard). Using Warren, Fey, and Yoder's (2007) recommended approach to assessing intervention dosage, we used the scoreboard data to ascertain the *cumulative intervention intensity* of the GBG for each class in a given year (put simply, the total number of minutes' exposure to the game).

Given the lack of agreed thresholds of implementation ratings for the GBG (meaning that a binary 'on/off treatment' classification would be inappropriate), we applied approaches adopted in a previous implementation-outcomes analysis of universal, school-based intervention (Humphrey, Barlow, & Lendrum, 2017). Accordingly, we used the above data to classify each class/teacher as 'low', 'moderate' or 'high' for each aspect of implementation using a distributional cut-point method (low,  $< -1$  SD; moderate,  $-1$  to  $+1$  SD; and high,  $> +1$  SD). Of note is the fact that these designations were statistical rather than qualitative (that is, they are based on relative position in the distribution as opposed to being based on arbitrary thresholds of 'good' implementation; Durlak & DuPre, 2008). The exception to this was reach, for which the proportion of pupils present was categorised as low ( $<90\%$ ), as moderate (90-99%) and high (100%).

The four dimensions of implementation (fidelity/quality, participant responsiveness, reach, and dosage) were subsequently modelled as explanatory variables at the class level in two-level (class, pupil) MLMs for each outcome measure (in subsequent dummy coding, 'low' was the designated reference group). Gender and FSM were fitted at the pupil level alongside baseline (T1) score, with the follow-up (T3) scores as the response variable. Due to the number of schools/classes that discontinued implementation, and the movement of pupils across classes (e.g. in some two/three and mixed form entry schools, class composition each year), analyses were conducted separately for the implementation data from the first and second years of the trial. As per our impact analyses, we began with complete cases (Model 1.1). Subsequently, we assessed the sensitivity of our findings to the modelling of implementation data as continuous variables in order to increase power (Durlak & DuPre, 2008) (Model 1.2). The sensitivity of these findings to the use of both fully *and* partially observed data via multiple imputation (MI) was then assessed (Models 2.1 and 2.2, respectively) using the procedures outlined above.

### Qualitative case studies

We conducted longitudinal case studies of a convenience sample of six GBG schools recruited at the initial training stage of the trial (although as noted, two of these schools ceased implementation after one year). The two main, inter-related purposes of the case studies were to: (1) develop our understanding of *how* the GBG was implemented and *why* it was implemented in this way; and (2) explain and add contextual detail to our quantitative findings.

In terms of *how* the GBG was implemented, we focused on the following dimensions: fidelity (e.g. to what extent do teachers adhere to the GBG guidance?), dosage (e.g. how frequently is the GBG played and for how long?), quality (e.g. how well do teachers deliver the components of the GBG?), participant responsiveness (e.g. to what extent do children engage with the GBG?), reach (e.g. what is the rate and scope of participation in the GBG across the class?), programme differentiation (e.g. to what extent can the GBG be distinguished from other, existing behaviour management practices?), and adaptations (e.g. what is the nature and extent of changes made to the GBG during the course of implementation?).

In terms of *why* it was implemented in this way, we used the case studies to explore a range of factors affecting implementation at different domains/levels: preplanning and foundations (e.g. buy-in), implementation support system (e.g. on-going external support), implementation environment (e.g. time constraints), implementer factors (e.g. experiences, skills and confidence in delivery), and programme characteristics (e.g. flexibility) (Durlak & DuPre, 2008; Forman, Olin, Hoagwood, & Crowe, 2009; Greenberg, Domitrovich, Graczyk, Zins, & Services, 2005).

Overarching the above was a social validity framework (Wolf, 1978) focusing on key tenets of acceptability, feasibility and utility (e.g. does the intervention meet schools' perceived needs? How well received is the intervention among staff and pupils? Can the intervention be delivered successfully?). As is the norm in case study research, we made use of a range of methods (e.g. interviews, focus groups, observations and document analysis), and informants (e.g., pupils, teachers, school leaders, parents, GBG coaches).

Two case study visits were conducted in each year of the trial (November/December and February/March/April) in order to explore schools' progression through the various phases of implementation. Thus, early visits focused primarily on pre-implementation issues (e.g. exploring foundations for the GBG and decisions to join the trial) and initial implementation. Over time, the focus shifted to continuing implementation, and perceptions of impact and sustainability. Table 1 summarises the case study school data collection.

**Table 1: Case study school data collection summary**

School	Visit 1 Autumn term 2015	Visit 2 Spring term 2016	Visit 3 Autumn term 2016	Visit 4 Spring term 2017
1	<ul style="list-style-type: none"> <li>• GBG lead interview</li> <li>• Year 3 teacher interviews</li> <li>• Informal lesson observation</li> <li>• Field notes</li> </ul>	<ul style="list-style-type: none"> <li>• Year 3 teacher interview</li> <li>• Pupil focus group</li> <li>• Parent interview</li> <li>• Formal lesson observation</li> <li>• Field notes</li> </ul>	<ul style="list-style-type: none"> <li>• Year 4 teacher interviews (x2)</li> <li>• Informal observation</li> <li>• Field notes</li> </ul>	<ul style="list-style-type: none"> <li>• Year 4 teacher interview</li> <li>• Pupil focus group</li> <li>• Parent interview</li> <li>• Formal lesson observation</li> <li>• Field notes</li> </ul>
2	<ul style="list-style-type: none"> <li>• GBG lead interview</li> <li>• Year 3 teacher interviews (x3) plus TA</li> <li>• Informal lesson observation (x3)</li> <li>• Field notes</li> </ul>	<ul style="list-style-type: none"> <li>• Formal lesson observation (X2)</li> <li>• Field notes</li> </ul>	Ceased implementation	
3	<ul style="list-style-type: none"> <li>• Year 3 teacher interviews</li> <li>• Informal lesson observation</li> <li>• Field notes</li> </ul>	<ul style="list-style-type: none"> <li>• Year 3 teacher interview</li> <li>• Pupil focus group</li> <li>• Formal lesson observation</li> <li>• Field notes</li> </ul>	<ul style="list-style-type: none"> <li>• Year 4 teacher interviews (x1)</li> <li>• Informal observation</li> <li>• Field notes</li> </ul>	<ul style="list-style-type: none"> <li>• Year 4 teacher interview</li> <li>• HT interview</li> <li>• Formal lesson observation</li> <li>• Field notes</li> </ul>
4	<ul style="list-style-type: none"> <li>• Head Teacher interview</li> <li>• Year 3 teacher interviews</li> <li>• Informal lesson observation</li> <li>• Field notes</li> </ul>	<ul style="list-style-type: none"> <li>• Year 3 teacher interview</li> <li>• Pupil focus group</li> <li>• Parent interview</li> <li>• Formal lesson observation</li> <li>• Field notes</li> </ul>	<ul style="list-style-type: none"> <li>• Head Teacher Interview</li> <li>• Year 4 teacher Interviews (x1)</li> <li>• Informal observation</li> <li>• Field notes</li> </ul>	<ul style="list-style-type: none"> <li>• Head Teacher interview</li> <li>• Year 4 teacher interviews (x1)</li> <li>• Pupil focus group</li> <li>• Formal lesson observation</li> <li>• Field notes</li> </ul>
5	<ul style="list-style-type: none"> <li>• GBG lead/Deputy Head Teacher Interview</li> <li>• Head Teacher interview</li> <li>• Year 3 teacher interviews</li> <li>• Informal lesson observation</li> <li>• Field notes</li> </ul>	<ul style="list-style-type: none"> <li>• Year 3 teacher interview</li> <li>• Pupil focus group</li> <li>• Formal lesson observation</li> <li>• Field notes</li> </ul>	Ceased implementation	
6	<ul style="list-style-type: none"> <li>• Head teacher interview</li> <li>• Year 3 teacher interviews (x2)</li> <li>• Informal lesson observation (x2)</li> <li>• Field notes</li> </ul>	<ul style="list-style-type: none"> <li>• Year 3 teacher interview (x2)</li> <li>• Pupil focus group (x2)</li> <li>• Parent interview (x2)</li> <li>• Formal lesson observation (x2)</li> <li>• Field notes</li> </ul>	<ul style="list-style-type: none"> <li>• Year 4 teacher interviews (x2)</li> <li>• Informal observation</li> <li>• Field notes</li> </ul>	<ul style="list-style-type: none"> <li>• Year 4 teacher interviews (x2)</li> <li>• Formal observation</li> <li>• Field notes</li> </ul>

Qualitative data were analysed thematically using the principles and process outlined by Braun & Clarke (2006) (e.g. familiarisation, generating initial codes, searching for themes, reviewing themes, defining

and naming themes, report production). We undertook a hybrid inductive-deductive approach to thematic identification. The deductive aspect drew upon key sources in the implementation science literature (e.g. Durlak and DuPre's (2008) review and model of implementation dimensions and the factors affecting implementation; Fixsen, Naoom, Blase, Friedman, & Wallace's (2005) conceptual framework of the stages of implementation) and our primary orienting concepts (e.g. acceptability, feasibility, utility), which informed the development of our thematic framework and/or interpretation of qualitative data. The inductive aspect drew directly from the data itself, thus enabling the identification of emergent themes during the process of analysis. Data were analysed both within- and across-case. We present the latter in the IPE chapter of this report. The within-case analyses are available on request from the evaluation team.

## Costs

The cost of GBG in the context of the trial reported here differs significantly from the 'real world' costs of the intervention. Delivery costs in the trial were part-funded by the EEF, with schools assigned to the intervention arm paying a nominal fee based upon their size (£750 per form-entry, so an 'average' two-form entry school paid £1,500). Included within this were:

- GBG materials and resources (manuals, charts, stamps, recording booklets and posters; one pack per class).
- Two training events in each year of the trial led by Mentor UK GBG coaches (a two-day initial training event in September, and one-day top-up training in January).
- On-going coaching support (this equated to an average of one school visit per month to support implementation – see TIDieR for details).

Thus, below we provide the typical costs for a school under normal circumstances. Cost data was provided by Mentor UK, based on a cohort of 10 schools signing up to the GBG and thus spreading centralized costs (e.g. administration team, programme director) between them.<sup>13</sup>

Costs were calculated based on the direct costs of the intervention, resources, and training in the first year, and subsequent annual running costs over a three year period of implementation, and assuming the participation of all Key Stage Two classes in single-form entry schools with approximately 27 pupils per class pupils (DfE, 2017).

Please see appendix 12 for how cost ratings are calculated.

---

<sup>13</sup> Currently a single school cost model is not feasible, as centralized costs (e.g. coaching team, administration team, programme director) would be prohibitive. Thus, 1 year running costs for a 2-form entry school would be over £100,000. Mentor's proposed 10-school model offers economies of scale.

## Timeline

The timeline for the project is outlined in Table 2 below.

**Table 2: GBG project timeline**

Date	Activity
January-June 2015	<b>Ethical approval process completed, schools recruited</b>
April-June	<b>NPD request for class lists to populate baseline surveys</b>
May-July 2015	<b>Baseline (T1) outcome data collection</b>
July 2015	<b>Randomisation</b>
September 2015	<b>GBG training events with Mentor UK GBG coaches (Year 3 teachers)</b>
September 2015	<b>GBG implementation begins (Year 3 classes)</b>
September 2015	<b>GBG case study schools recruited</b>
November-December 2015	<b>Case study fieldwork visit 1</b>
December 2015	<b>NPD request for baseline (T1) academic and pupil demographic data</b>
January 2016	<b>GBG follow-up training events with Mentor UK GBG coaches (Year 3 teachers)</b>
January-March 2016	<b>GBG lesson observations (Year 3 classes at GBG schools)</b>
March-April 2016	<b>Case study fieldwork visit 2</b>
May-July 2016	<b>Interim (T2) outcome data collection</b>
September 2016	<b>GBG training events with Mentor UK GBG coaches (Year 4 teachers)</b>
September 2016	<b>GBG implementation begins (Year 4 classes)</b>
November-December 2015	<b>Case study fieldwork visit 3</b>
January 2017	<b>GBG follow-up training events with Mentor UK GBG coaches (Year 4 teachers)</b>
January-March 2017	<b>GBG lesson observations (Year 4 classes)</b>
March-April 2017	<b>Case study fieldwork visit 4</b>
May-July 2017	<b>Post-intervention (T3) outcomes data collection</b>
August-October 2017	<b>Data cleaned, screening and analyses, report write up.</b>

## Impact evaluation

### Participants

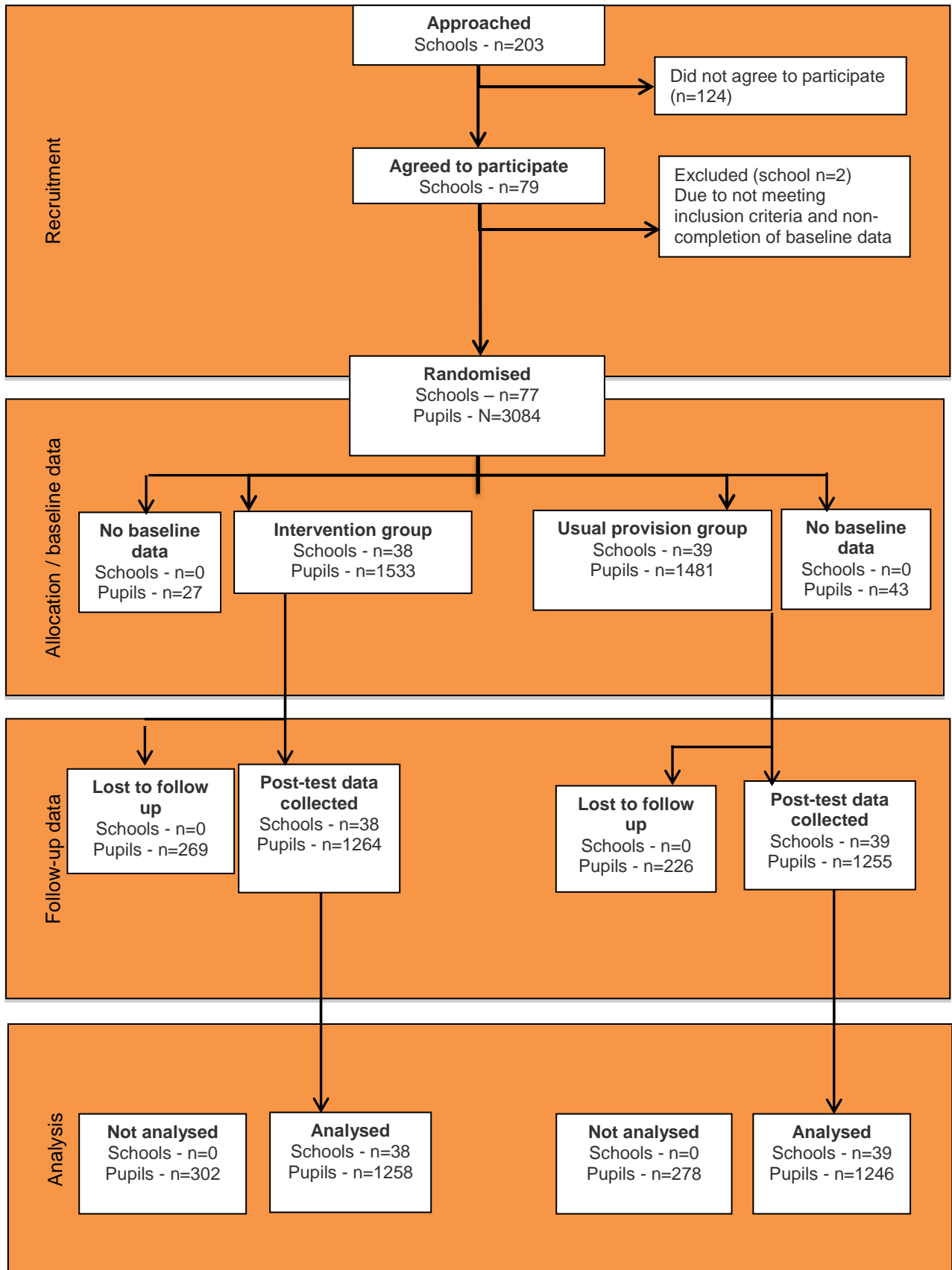
As noted earlier, 77 schools (N=3,084 pupils) were recruited, of whom 38 were randomly allocated to the GBG arm and 39 to the usual provision arm. In terms of the primary outcome, Key Stage 1 reading points at baseline (T1) were available for 3,014 pupils (98%). All missing T1 cases were due to pupils being opted out or the lack of a match in the NPD. Post-intervention (T3) HGRT data were missing for 565 (18.3%) pupils, in cases where they had left the school (n=390, 12.6%) or were absent on day of testing (n=175, 5.7%). Complete data (e.g. both T1 and T3) were available for n=2,504 pupils (81%). See Figure 2 for participant flow. The resultant sample size meant that our complete case analysis (Model 1.1) was powered for a MDES of 0.15 (see Table 3 below).

**Table 3: Minimum detectable effect size at different stages of the trial**

Stage	N [schools/pupils] (n=intervention; n=usual provision)	Correlation between pre-test and post-test	ICC	Power	Alpha	MDES
<b>Protocol</b>	72/2880 (36/1440; 36/1440)	-	0.06	80%	0.05	0.20
<b>Randomisation</b>	77/3014 (38/1533; 39/1481)	0.70	0.06	80%	0.05	0.13
<b>Analysis</b>	77/2504 (38/1258; 39/1246)	0.74	0.08	80%	0.05	0.15



Figure 2: Participant flow diagram



## Pupil characteristics

### Imbalance at baseline

Baseline demographic, attainment, school-level and other relevant characteristics of the final sample of 77 schools and N=3,084 pupils are presented below (Table 4). The composition of the trial school sample mirrored that of primary schools in England in respect of size and the proportion of students speaking EAL, but trial schools contained significantly larger proportions of children with SEND and eligible for FSM, in addition to lower rates of absence and attainment (assessed via single sample t-tests; see Appendix 7). There were no significant differences between trial arms for any of these variables ( $F(7, 68) = 0.78, p = .608$ ). Eight of the 77 schools (10%) were rated “Outstanding” by Ofsted, 54 (70%) as “Good”, nine (12%) as “Requires Improvement”, and six (8%) as “Inadequate”.

ES differences between pupil-level outcome variables at baseline were very small (KS1 reading points ES = -0.11; concentration problems ES = 0.01; disruptive behaviour ES = 0.11; pro-social behaviour ES = -0.03). Thus, balance on key observables in the sample analysed was considered to be good.

**Table 4: Baseline comparison of school and pupil characteristics.**

Variable	Intervention group (N=38)		Usual provision group (N=39)	
	n (missing)	Mean (SD)	n (missing)	Mean (SD)
<b>School-level (continuous)</b>				
Number of full-time equivalent (FTE) pupils on roll	38 (0)	298.21 (134.33)	39 (0)	315.41 (186.65)
Attendance – overall absence (% half days)	38 (0)	4.26 (0.90)	39 (0)	4.17 (0.96)
Proportion eligible for FSM	38 (0)	27.56 (13.37)	39 (0)	24.46 (13.30)
Proportion speaking EAL	38 (0)	22.01 (26.05)	39 (0)	23.19 (27.91)
Proportion with SEND	38 (0)	20.85 (9.30)	39 (0)	18.17 (5.94)
Proportion achieving level 4+ in English and maths	38 (0)	76.21 (12.05)	39 (0)	74.87 (10.96)
<b>Pupil-level (categorical)</b>				
	n (missing)	Percentage	n (missing)	Percentage
Proportion of male pupils	1559 (0)	50.4%	1525 (0)	54.9%
Proportion eligible for FSM	1543 (16)	27.4%	1493 (32)	22.8%
Proportion speaking EAL	1543 (16)	26.1%	1493 (32)	29.5%
Proportion with SEND	1543 (16)	23.1%	1493 (32)	18%
Proportion scoring in at-risk range for SDQ conduct problems	1498 (61)	13.2%	1471 (54)	17.9%
<b>Pupil-level (continuous)</b>				
	n (missing)	Mean (SD)	n (missing)	Mean (SD)
KS1 reading points (1-53)	1533 (27)	15.61 (3.99)	1481 (43)	16.06 (4.00)
Concentration problems (1-6)	1498 (61)	2.60 (1.13)	1469 (56)	2.55 (1.15)
Disruptive behaviour (1-6)	1497 (62)	1.71 (0.81)	1469 (56)	1.61 (0.81)
Pro-social behaviour (1-6)	1498 (61)	4.89 (0.88)	1469 (56)	4.94 (0.92)

### Missing data

Missing data is common in educational research, resulting in a loss of valuable information and potential selection bias (Pampaka, Hutcheson & Williams, 2016). It is therefore becoming standard practice to (a) report the scale of missing data, (b) perform analyses to investigate correlates of missingness, and (c) utilise approaches that can deal with incomplete datasets (such as multiple imputation), thereby minimising the bias associated with attrition. As reported above, complete data were available for n=2,504 (81%) of the sample, leaving partially observed data for n=580 (19%), of whom 302 (10%) were from GBG schools and 278 (9%) from usual provision schools. Following guidance produced by Pampaka, Hutcheson, and Williams, (2016), missingness was investigated using logistic regression in

MLwiN2.36. Missingness at T3 (0=no, 1=yes) was used as the response variable, with other study data as explanatory variables (e.g. KS1 reading points, SDQ conduct problems and TOCA scores at baseline, gender, and FSM at the child level, and trial group allocation at the school level). The resultant model (see Appendix 8) was statistically significant but the various explanatory variables yielded only marginal predictive power. Significant predictors of T3 missingness were lower KS1 reading points ( $\beta=-0.016$ ,  $p<.001$ ) and pro-social behaviour ( $\beta=-0.026$ ,  $p=.030$ ), and higher concentration problems at baseline ( $\beta=-0.019$ ,  $p=.035$ ). Accordingly, MI procedures were carried out in REALCOM-Impute, using the missing at random assumption (Carpenter, Goldstein, & Kenward, 2011), wherein missingness is considered to be conditional on other observed variables. Demographic variables (e.g. gender, FSM eligibility, ethnicity, EAL, SEND provision), explanatory outcome variables (e.g. KS1 reading and TOCA scores), and the constant were entered as auxiliary variables and used to impute missing values. REALCOM-Impute default settings of 1000 iterations and a burn-in of 100, refresh of 10, were used, following guidance for multi-level imputation with mixed response types (Carpenter et al., 2011).

## Outcomes and analysis

Table 5 below provides basic descriptive statistics for our pupil-level outcome variables at baseline (T1) and post-intervention follow-up (T3). These data do not appear to be indicative of an intervention effect for any outcome.

**Table 5: Mean (SD) pupil outcomes at baseline (T1) and post-intervention (T3) follow-up**

	Intervention group		Usual provision group	
	Baseline	Follow-up	Baseline	Follow-up
<b>KS1 reading points (1-53)</b>	15.61 (3.99)	-	16.06 (4.00)	-
<b>HGRT raw score (0-53)</b>	-	32.49 (0.29)	-	33.05 (0.29)
<b>Concentration problems (1-6)</b>	2.60 (1.13)	2.55 (1.13)	2.55 (1.15)	2.50 (1.13)
<b>Disruptive behaviour (1-6)</b>	1.71 (0.81)	1.74 (0.86)	1.61 (0.81)	1.65 (0.84)
<b>Pro-social behaviour (1-6)</b>	4.89 (0.88)	4.81 (0.93)	4.94 (0.92)	4.93 (0.95)

Table 6 below summarises our findings for Model 1.1 (complete case analysis).

**Table 6: Intention to treat and sub-group analyses: pupil-level outcomes**

Outcome	Intervention group		Usual provision group		n in model (intervention; usual provision)	Hedges <i>g</i> (95% CI)	<i>p</i>
	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)			
<b>H1: Main effect of intervention – ITT</b>							
Reading	1264 (296)	32.49 (31.92-33.06)	1255 (269)	33.05 (32.47-33.63)	2504 (1258; 1246)	0.03 (-0.08-0.16)	.299
Concentration problems	1202 (358)	2.55 (2.48-2.61)	1310 (214)	2.50 (2.43-2.56)	2469 (1176; 1293)	0.03 (-0.15-0.21)	.364
Disruptive behaviour	1202 (358)	1.74 (1.69-1.79)	1310 (214)	1.64 (1.60-1.69)	2469 (1176; 1293)	0.06 (-0.09-0.22)	.219
Pro-social behaviour	1203 (357)	4.81 (4.77-4.86)	1310 (214)	4.93 (4.88-4.98)	2469 (1176; 1293)	-.13 (-0.36- -0.11)	.135
<b>H2: Interaction effects - boys at risk of conduct problems</b>							
Reading	137 (41)	29.46 (27.64-31.38)	118 (41)	29.22 (27.11-31.33)	2466 (1237; 1229)	0.05 (-0.28-0.38)	.394
Concentration problems	135 (43)	3.42 (3.23-3.60)	133 (26)	3.75 (3.57-3.93)	2468 (1176;1292)	-0.29 (-0.66-0.08)	.063
Disruptive behaviour	135 (43)	2.72 (2.54-2.90)	133 (26)	2.95 (2.75-3.14)	2468 (1176;1292)	-0.30 (-0.68-0.07)	.053
Pro-social behaviour	135 (43)	4.21 (4.04-4.38)	133 (26)	3.94 (3.77-4.11)	2468 (1176;1292)	0.12 (-0.31-0.55)	.298
<b>H3: Interaction effects - pupils eligible for FSM</b>							
Reading	327 (96)	29.52 (28.41-30.62)	264 (77)	30.03 (28.74-31.33)	2504 (1258; 1246)	0.05 (-0.07-0.18)	.215
Concentration problems	328 (95)	2.88 (2.76-3.01)	284 (57)	2.82 (2.68-2.96)	2463 (1174; 1289)	-0.02 (-0.16-0.13)	.397
Disruptive behaviour	328 (95)	1.91 (1.80-2.02)	284 (57)	1.82 (1.75-1.93)	2463 (1174; 1289)	0.09 (-0.06-0.23)	.125
Pro-social behaviour	328 (95)	4.63 (4.53-4.73)	284 (57)	4.70 (4.58-4.81)	2463 (1174; 1289)	0.10 (-0.07-0.26)	.131

**H1: Children in primary schools implementing the GBG over a two-year period will demonstrate significant improvements in reading (1a), concentration problems (1b), disruptive behaviour (1c) and pro-social behaviour (1d) when compared to those children attending usual provision schools.**

There was no significant impact of the GBG at the ITT level on children's reading (ES = 0.03, CI = -0.08 to 0.16), concentration problems (ES = 0.03, CI = -0.15 to 0.21), disruptive behaviour (ES = 0.06, CI = -0.09 to 0.22) or pro-social behaviour (ES = -0.13, CI = -0.36 to 0.11). Full models, along with accompanying sensitivity (Models 1.2 to 1.5) and MI analyses (Models 2.1 to 2.5), are provided in Appendix 9. The findings of the complete case analyses were borne out in all of these models; put another way, our findings were not sensitive to any changes we made to our modelling parameters noted in the 'Analysis' section of the Methods chapter (e.g. the inclusion of minimisation variables at the school level; MI of missing data).

**H2: The effects outlined in H1 above will be amplified for boys exhibiting borderline/abnormal levels of conduct problems at baseline.**

There was no significant differential impact of the GBG for the 'at-risk boys' subgroup in relation to the primary outcome, reading (ES = 0.05, CI = -0.28 to 0.38). There were, however, marginal non-significant trends (i.e.  $p < .10$ ) indicative of favourable intervention effects among this subgroup for both concentration problems (ES = -0.29, CI = -0.66 to 0.08) and disruptive behaviour (ES = -0.30, CI = -0.68 to 0.07). There was no significant impact of the GBG for our 'at-risk boys' subgroup in relation to pro-social behaviour (ES = 0.12, CI = -0.31 to 0.55). Full models, along with accompanying sensitivity (Models 1.2 to 1.5) and MI analyses (Models 2.1 to 2.5), are provided in Appendix 9. As above, our findings were not generally sensitive to any changes we made to our modelling parameters (e.g. the inclusion of minimisation variables at the school level; MI of missing data), with the following exceptions: (i) the marginal, non-significant trend observed in relation to concentration problems (Model 1.3) was no longer evident in the corresponding MI analysis (Model 2.3); and (ii) the marginal, non-significant trend observed in relation to disruptive behaviour became statistically significant in both the complete case (Model 1.5) and MI version (Model 2.5) of the analysis in which all explanatory variables were included simultaneously.

**H3: The effects outlined in H1 above will be amplified for children eligible for FSM.**

There was no significant differential impact of the GBG for the FSM subgroup in relation to reading (ES = 0.05, CI = -0.07 to 0.18), concentration problems (ES = -0.02, CI = -0.16 to 0.13), disruptive behaviour (ES = 0.09, CI = -0.06 to 0.23) or pro-social behaviour (ES = 0.10, CI = -0.07 to 0.26). Full models, along with accompanying sensitivity (Models 1.2 to 1.5) and MI analyses (Models 2.1 to 2.5), are provided in Appendix 9. As above, our findings were not generally sensitive to any changes we made to our modelling parameters, with the following exception: a statistically significant intervention effect was found in relation to reading in the complete case version of the analysis in which all explanatory variables were included simultaneously (Model 1.5); in the MI version of this analysis (Model 2.5), this attenuated to a marginal, non-significant trend.

**H5: Teachers implementing the GBG will demonstrate measurable improvements in self-efficacy in classroom management (5a), classroom stress (5b), and retention (5c), when compared to teachers in usual provision schools.**

Single level linear regression models using complete cases, summarised in Table 7 below, showed no significant impact of the GBG on teachers' self-efficacy in classroom management ( $\beta = 0.11$ , ES = 0.06, CI = -0.19 to 0.31), classroom stress ( $\beta = -0.07$ , ES = -0.05, CI = -0.30 to 0.20), or retention ( $\beta = -0.02$ , ES = -0.01, CI = -0.26 to 0.24). Subsequent models including both complete and partially observed cases to account for missing data (15.1-16.8%) led to the same conclusion (see Appendix 10).

**Table 7: Intention to treat analyses: teacher-level outcomes**

Outcome	Intervention group		Usual provision group		n in model (intervention; usual provision)	Hedges g (95% CI)	p
	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)			
<b>Classroom management (1-9)</b>	111 (29)	8.18 (8.00-8.37)	119 (20)	8.09 (7.94-8.24)	198 (91; 107)	0.06 (-0.19-0.31)	.345
<b>Stress (0-4)</b>	111 (29)	1.58 (1.43-1.73)	119 (20)	1.61 (1.48-1.75)	195 (90; 105)	-0.05 (-0.30-0.20)	.439
<b>Retention (1-6)</b>	111 (29)	3.11 (2.84-3.37)	119 (20)	3.29 (3.06-3.52)	198 (91; 107)	-0.01 (-0.26-0.24)	.889

## Costs

As noted earlier, the number of schools taking part in the GBG in a given area makes a considerable difference to the interventions costs as there are some fixed centralised costs and economies of scale. The cost of support from AIR is an example of where economies of scale are material. Additionally, the length of a programme and the average number of pupils per class are clearly variable factors. At the time of writing, Mentor UK have budgeted for the GBG to be rolled out in a single local authority involving a consortium of 10 schools over one financial year.

The estimated initial start-up cost per school is £4,000, **£37.04 per pupil**. Over three years there would be some savings (e.g. initial teacher training, and the estimated costs would be fairly close to pro rata, so c. £11,000 per school/three years, **£33.95 per pupil**, see Table 8). If the number of schools in a programme were much higher, then costs would be lower. For example Mentor UK have also budgeted for a 60-school programme with initial start-up cost of £3,700 per school pa, £35.53 per pupil, per annum.

**Table 8: Start-up and running costs of the GBG**

10 school consortium example	
<b>Cost per school, per annum</b>	£4,000
<b>Cost per pupil, per annum</b>	£37.04
<b>Cost per school over 3 years</b>	Start-up cost - £4, 000 Running cost per annum - £3,500 Total - £11,000
<b>Cost per pupil over 3 years</b>	£33.95

As the GBG is played during a typical lesson/activity, there is minimal additional teaching time or staffing requirements outside of normal practice. Teachers are required to attend an initial two-day training event, held in September, and a further one-day follow-up training event, held in January – approximately 21 hours of training in total in the first year. Additional supply cover or reallocation of staff may be required to cover these three days of training.

Preparation time is estimated to be marginal and would typically involve the organisation of pupils into teams, allocation of team leaders, and maintaining the posters and resources. There are also monthly GBG coach visits, which typically involve an observation of the game followed by a meeting of up to 30 minutes for discussion. This equates to approximately 10 visits per school year (October-July), involving up to 5 hours of coaching support meetings. If any of these meetings are conducted during teaching time, reallocation of staff would be required to cover the teacher's class.



## Implementation and process evaluation

### Usual practice

Descriptive statistics for each domain of the usual practice survey are shown in Table 9 below. Of particular note is that the use of reward systems to manage pupil behaviour – a central feature of the GBG - was commonplace at baseline across both arms of the trial. A 2 x 2 x 4 (*group*: GBG vs usual practice; *time*: baseline vs follow-up; *domain*: general behaviour management, use of reward systems, systems and procedures for managing disruption, use of physical and verbal reprimands) mixed analysis of variance (ANOVA) conducted on the usual practice survey data revealed:

- A main effect of time ( $F(1, 113) = 15.47, p < .001$ )
- No main effect of trial group ( $F(4, 162) = 1.47, p = .213$ )
- A main effect of domain ( $F(3, 339) = 126.56, p < .001$ )
- No interaction between group and time ( $F(1, 113) = 0.06, p = .813$ )
- A significant interaction between time and domain ( $F(3, 339) = 7.93, p < .001$ )
- No interaction between group, time and domain ( $F(3, 339) = 0.11, p = .922$ )

Thus, teachers reported increased use of a range of behaviour management practices from baseline to follow-up (main effect of time), and specifically the use of reward systems (interaction between time and domain). However, this did *not* vary as a function of trial group (no interaction between group, time and domain; that is, teachers in *both* trial arms reported an increase in the use of reward systems over time).

**Table 9: Teachers' self-reported usual practice in behaviour management**

		GBG		Usual provision	
		Baseline Mean (SD)	Follow-up Mean (SD)	Baseline Mean (SD)	Follow-up Mean (SD)
<b>Year 3 teachers (2015/16, n=135)</b>	General behaviour management (0-20)	17.25 (1.84)	17.75 (1.61)	17.26 (2.09)	17.72 (1.73)
	Use of reward systems (0-24)	10.63 (3.58)	13.85 (4.27)	10.93 (3.64)	13.96 (4.09)
	Systems and procedures for managing disruption (0-15)	10.55 (2.83)	11.22 (2.76)	10.14 (3.02)	10.70 (2.97)
	Use of physical and verbal reprimands (0-27)	14.24 (5.19)	14.02 (4.91)	13.05 (5.00)	13.47 (4.93)
<b>Year 4 teachers (2016/17, n=144)</b>	General behaviour management (0-20)	17.26 (2.13)	17.56 (1.95)	17.28 (1.99)	17.91 (1.88)
	Reward systems (0-24)	12.15 (4.17)	13.75 (4.17)	11.67 (3.78)	13.13 (5.01)
	Systems and procedures for managing disruption (0-15)	10.65 (2.59)	10.25 (2.81)	10.58 (3.17)	10.00 (3.09)
	Use of physical and verbal reprimands (0-27)	12.82 (5.14)	14.04 (5.90)	12.96 (4.69)	13.43 (5.09)
<b>All teachers (2015-17, n=279)</b>	General behaviour management (0-20)	17.26 (1.97)	17.65 (1.78)	17.27 (2.04)	17.82 (1.81)
	Reward systems (0-24)	11.37 (3.98)	13.80 (4.20)	11.30 (3.71)	13.55 (4.67)
	Systems and procedures for managing disruption (0-15)	10.60 (2.70)	10.72 (2.81)	10.36 (3.09)	10.35 (3.04)
	Use of physical and verbal reprimands (0-27)	13.56 (5.18)	14.03 (5.38)	13.01 (4.83)	13.45 (4.99)

Further analysis of individual items in the above survey that are reflective of key elements of the GBG proved to be particularly instructive in terms of establishing programme differentiation and the counter-factual. Consider the following, taken from the baseline data:

- "I establish and maintain a set of classroom rules" (94.0% GBG; 95.1% usual provision, endorsed 'yes').
- "I communicate clear expectations about rules and pupils' responsibilities, e.g. through posters" (89.6% GBG; 90.2% usual provision, endorsed 'yes').

- “I observe and monitor pupils’ behaviour in the classroom” (97.4% GBG; 100% usual provision, endorsed ‘yes’).
- “I use prizes as rewards for good behaviour” (53.4% GBG; 59.9% usual provision, endorsed ‘weekly’ or ‘every day’).
- “I use group rewards” (69.8% GBG; 66.6% usual provision, endorsed ‘weekly’ or ‘every day’).

In summary, our usual practice survey data demonstrated balance between the trial arms in relation to use of behaviour management strategies. However, the data also appeared to be indicative of relatively low levels of ‘programme differentiation’ (e.g. the extent to which an intervention can be distinguished from existing practice; Durlak & DuPre, 2008) in relation to the GBG. At baseline, teachers reported using a number of behaviour management strategies that are core to the GBG as part of their usual practice.

## Implementation of the GBG

### Discontinuation of implementation

As noted earlier, nine of the 38 GBG schools (24%) ceased implementation before the conclusion of the trial. The following analysis is derived from ‘exit interviews’ conducted with two schools, in addition to email communication between the other schools and the project delivery team (Mentor UK).

A principal factor was the extent to which key members of the school staff accepted the GBG. It was typically members of senior leadership team (SLT) that made the decision to join the trial: “*I was sent the email about whether or not we wanted to take part in the Good Behaviour Game... I then discussed it with the Head and she agreed to it because I was so enthusiastic*” (school 7, SENCO). However, teachers did not necessarily share their enthusiasm: “*I think where we failed was to sell it to the two teachers that were going to end up having to deliver it... I tried to share my enthusiasm for it but I don’t think they completely understood what they were being let into*” (school 7, SENCO). Teachers’ views of the GBG appeared to be underpinned by the relationship between the amount of time and effort invested in implementation (feasibility) and the perceived changes in outcomes (utility):

*“I found that every single time... I was getting quite frustrated with how the game was working... I was waiting for it to get better, and it wasn’t getting better”* (school 8, teacher F)

*“[the implementing teacher] wasn’t seeing any benefit for what felt like quite a lot of effort, so although she was doing it really well she wasn’t sold on the long-term picture”* (school 7, SENCO)

On reflection, members of SLT noted the importance of ensuring that teachers fully understood and agreed to the implementation commitment, as without this there was evidence of conflict with other priorities: “*my two members of staff didn’t know what they were agreeing to, they didn’t know how much work was involved and I think that needed to be made clear... there may well be quite a lot of lesson time given over to playing the game... while the children and the teachers get used to it*” (school 7, SENCO).

The amount of lesson time required to deliver the GBG was a factor that made teachers reluctant to continue implementation: “*it took too long to implement based on the beginning parts and the end parts*” (school 8, teacher F). Thus, it was felt that the intervention was in conflict with the pressure to deliver the academic curriculum, and that “*maybe the people that designed it hadn’t designed it with necessarily... [with] the British curriculum in mind, there’s too many things for us to have to cram into a year*” (school 8, teacher F), as “*that’s time out of curriculum and these days schools can be quite sensitive about that*” (school 7, SENCO). This suggests that teachers who perceived the GBG competing with curriculum time, rather than complementing it were less likely to continue with

implementation: *“I think for a simple life there's so much going on in a school [the Head Teacher] allowed [the teachers] to withdraw from it”* (school 7, SENCO).

A further factor that appeared to influence decisions to discontinue implementation was the discordance between the underlying principles of GBG and teachers' preferred pedagogical and classroom management approaches. Specifically, teachers struggled with the concept of not interacting with children while the game was in play, as *“you cannot give them something that's new because you cannot help them, but as a teacher that's what I'm there for...I'm there to help the students so it really frustrated me that I couldn't work with any of them”* (school 8, teacher F). As this lack of interaction is an essential component of the GBG, it was difficult for coaches and teachers to compromise on a solution that would satisfy both parties, and this led to further disengagement:

*“I raised this concern with the Good Behaviour Game people, they suggested that they come in and show us a class but then the lessons they were showing us... I just felt like it was a bit of a cop out... I felt like all of the games that we did do ended up being the type of things that you would do as a supply teacher with a class that you didn't know very well, easy things that you knew weren't going to go wrong. Then you're not challenging your students if you have to constantly pick subjects where you know nothing's going wrong”* (school 8, teacher F)

Although in some instances the GBG was perceived to have helped improve behaviour in the classroom, the lack of interaction with children was seen to be impeding the extent to which the teacher could aid their academic progression:

*“Their behaviour was beautiful, you know, they sat there, they got their stickers, every table won... However, the issue that I had was the lesson that I observed [there were a] handful of students who were the students that you would target, the ones that you would be sat with in a lesson basically saying ‘no that's not good enough do better do better.’”* (school 8, teacher F)

Problems integrating the GBG into existing systems and processes was another commonly cited concern among schools that discontinued implementation. For example, one Head Teacher noted that it was in *“conflict with the school's behaviour management policy”* (school 9, email communication from Mentor UK), while another school was concerned about the GBG, *“not being part of a whole-school initiative and the school's plans for this group of children going into Year four”* (school 5, email communication from Mentor UK). In a similar vein, one teacher *“felt reluctant and unable to integrate her chosen approach to classroom management with the game, particularly the aspect of non-interaction when the game is being played”* (school 10, email communication from GBG coach).

One school did withdraw from implementing the GBG early in the first year of the trial, claiming they had *“hoped they would be in the control group because of their profile... the school didn't particularly want to target a year three class”* (school 10, email communication from GBG coach). School staff did attend the initial training but reported they, *“came away with negative perceptions and unanswered questions about when and how to play the game”* and were also concerned *“her classroom organisation and management was being judged by her coach and felt uncomfortable about this”* (school 10, email communication from GBG coach).

Finally, several schools cited external factors such as staff turnover and changes within the school's structure, that meant they were *“not in a stable position”* (school 11, email communication from Head Teacher), due to losing key staff involved in implementation and therefore had *“a lot on their plate”* (school 11, email communication from school GBG coach). In another school, several members of staff had left and been replaced by newly qualified teachers (NQTs), who were finding it difficult to cope with the additional demands of GBG implementation:

*“It is proving increasingly difficult to maintain the expectations for training and visits. I am losing another member of staff this term to be replaced possibly by another NQT... The team needs to be allowed to work on... getting it right for the children. They are finding it difficult to meet the pressures of teaching and find the GBG is placing more strain on them as teachers and on the school in general, finding available staffing to cover for training and meeting purposes etc.” (school 2, email communication from Head Teacher)*

Another two schools cited impending academisation that meant they wanted to, *“focus on all the changes it will bring to move things forward... they would like to concentrate on a new whole school initiative for behaviour management strategy, so feel GBG could be an obstacle”* (school 12, email communication from Mentor UK), and that this process was leading to *“stress on the teachers”* (school 13, email communication from Mentor UK). Finally one school was keen to make the point that their decision to discontinue implementation had nothing to do with the intervention itself, and had noted it had been quite successful, but it was a result of difficult circumstances:

*“The head stressed that the decision to leave the GBG was not made against the intervention or the support that was offered to the school.”* (school 9, email communication from Mentor UK)

### Descriptive implementation data

We begin with descriptive data on GBG implementation through the course of the trial, derived from the structured observations and online scoreboard. These data, presented in Table 10, provide the mean scores and standard deviations in each year of the trial for fidelity/quality, participant responsiveness, reach (all expressed as a percentage), and dosage (reported as both the average number of minutes played per week and the average number of games played per week). Data was missing where classes had ceased implementation (2015/16 n= 5 classes; 2016/17 n=14 classes), where an observation did not take place (2015/16 n=1; 2016/17 n=0), or where a class teacher did not use the online scoreboard to record game data (2015/16 n=6; 2016/17 n=0).

**Table 10: Descriptive statistics on GBG implementation dimensions**

	2015/16		2016/17	
	No. classes	Mean (SD)	No. classes	Mean (SD)
<b>Fidelity/quality (%)</b>	54	69.79 (12.35)	45	70.11 (11.13)
<b>Participant responsiveness (%)</b>	51	74.51 (18.80)	43	69.07 (16.88)
<b>Reach (%)</b>	53	95.26 (7.79)	46	95.98 (6.23)
<b>Dosage (minutes per week)</b>	49	26.96 (17.61)	46	22.67 (16.97)
<b>Dosage (games per week)</b>	49	1.93 (1.15)	46	1.55 (0.94)

Thus, teachers played the GBG approximately twice a week in the first year of the trial, and between once and twice a week in the second year. The average game session length in both years was approximately 15 minutes. Fidelity/quality was relatively high in both years, indicating that teachers followed most of the prescribed procedures associated with the game (see section 5 of the intervention template in the Introduction chapter) and did so in an enthusiastic, engaging manner. Almost all children in a given class were present when the GBG was played, and they responded favourably (e.g. correcting their behaviour following an infraction), albeit with a drop in responsiveness in the second year of the trial (as shown in Table 10).

The online scoreboard data also enabled us to document temporal trends within each year with regard to frequency and duration of game playing. This is important given the expectation that both of these dosage dimensions develop over the course of the school year (e.g. in terms of frequency, from three times a week to every day; in terms of duration, from 10 minutes to a whole lesson). However, our data suggested that this was not the case. As can be seen in Table 11, teachers in both years of the trial very quickly settled into a routine with respect to their gameplay behaviour. After accounting for expected reductions in activity in certain months (e.g. in December and April, because of Christmas and Easter vacations, respectively), there was relative stability across the year in both 2015/16 and 2016/17 in terms of game play frequency. Similarly, game duration was consistently around 15 minutes and did not appear to increase over the course of the school year. Finally, there was also relative stability in terms of the number and duration of probe sessions.

**Table 11: Frequency and duration of games and probes in GBG classes<sup>14</sup>**

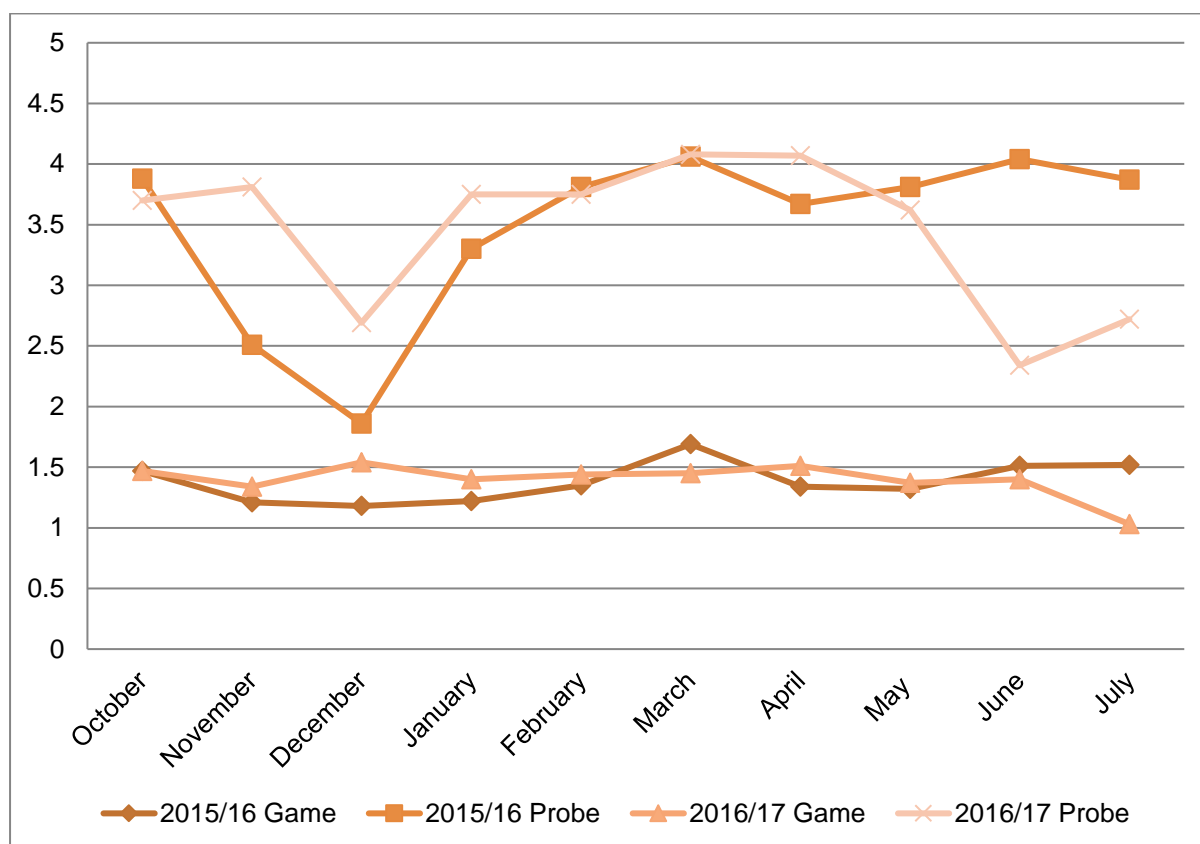
		Mean number of games	Mean game duration (minutes)	Mean number of probes	Mean probe duration (minutes)
<b>2015/16</b>	October	2.38	11.79	1.00	9.23
	November	3.93	11.17	1.00	10.00
	December	1.96	12.76	1.14	11.38
	January	6.73	13.49	2.32	11.47
	February	5.40	15.99	2.00	10.10
	March	6.55	14.51	2.32	11.98
	April	5.00	15.14	1.57	12.00
	May	6.35	15.01	1.81	12.53
	June	5.75	15.18	1.50	12.00
	July	2.03	13.37	1.36	12.11
<b>2016/17</b>	October	4.86	13.54	1.57	9.25
	November	6.18	14.18	2.55	9.04
	December	3.77	15.26	2.25	9.67
	January	6.27	13.27	2.58	10.84
	February	5.27	14.92	2.74	10.88
	March	6.34	15.43	2.79	11.85
	April	3.52	15.25	2.00	11.69
	May	6.24	15.17	2.33	13.00
	June	7.30	16.60	1.93	13.45
	July	2.94	18.48	2.00	11.82

Figure 3 displays the average number of infractions per team in both game and probe sessions. As a reminder, probe sessions involve the class teacher *covertly* recording data during an ordinary task/activity, following the same procedures as in a game session (e.g. the teacher monitors rule infractions among teams) but without explicitly setting up the rules and announcing infractions. In game sessions, the number of infractions is consistently around 1.5, meaning that teams are typically winning (e.g. four or fewer infractions). In terms of probe sessions, the number of infractions is higher (as would be expected) at around 3.5, but there is no evidence of generalisation occurring over time (e.g. a reduction in probe session infractions through the course of a school year).

<sup>14</sup> NB: The number of teachers/classes recording data for both game and probes sessions varies across the different months of the trial.



**Figure 3: Average number of rule infractions per team in game and probe sessions across the school year**



**H4: Variation in implementation fidelity/quality (4a), dosage (4b), reach (4c), and participant responsiveness (4d), will be significantly associated with reading and behavioural outcomes among pupils in schools implementing the GBG.**

Table 12 provides the descriptive statistics for the different dimensions of implementation in the low, moderate and high groups created using the distributional cut-point method noted earlier. Pupils in classes that had ceased implementation were removed from the analyses. The dataset for 2015/16 therefore consisted of 1,380 pupils across 55 implementing classes at 35 GBG schools. The dataset for 2016/17 consisted of 1,127 pupils across 46 implementing classes at 30 GBG schools.

**Table 12: Implementation group descriptive statistics (n, means, SDs)**

	2015/16			2016/17		
	Low	Moderate	High	Low	Moderate	High
<b>Procedural fidelity and quality (%)</b>	8/48.81 (7.27)	38/70.70 (7.18)	8/86.41 (2.72)	7/51.14 (5.42)	32/71.58 (6.74)	6/84.40 (3.17)
<b>Participant responsiveness (%)</b>	10/46.00 (9.66)	33/76.97 (10.15)	8/100.00 (0.00)	12/48.33 (3.89)	22/70.45 (7.22)	9/93.33 (5.00)
<b>Reach (%)</b>	11/84.48 (5.53)	9/94.69 (1.55)	33/100.00 (0.00)	5/80.92 (5.58)	18/95.03 (2.40)	23/100.00 (0.00)
<b>Dosage (total minutes played)</b>	10/139.00 (26.61)	30/497.17 (197.51)	9/1123.78 (104.47)	1/65.00 (0.00)	40/525.00 (240.57)	5/1834.20 (407.08)



Table 13 displays the descriptive statistics for primary and secondary pupil-level outcome measures in GBG classes at baseline and follow-up.

**Fidelity/quality (H4a/b):** Findings pertaining to the relationship between levels of fidelity/quality and pupil-level outcomes were mixed. Our analyses of the implementation data for the first year of the trial (2015/16) indicated that:

- Compared to low, and contrary to expectations, moderate levels of fidelity/quality were associated with significantly *lower* reading scores at follow-up (ES = -0.30, CI = -0.54 to -0.06).
- Compared to low, moderate and high levels of fidelity/quality were associated with significantly *higher* pro-social behaviour scores at follow-up (moderate ES = 0.40, CI = -0.05 to 0.85; high ES = 0.51, CI = -0.02 to 0.96).

Our analyses of the implementation data for the second year of the trial (2016/17) indicated that:

- Compared to low, and contrary to expectations, moderate levels of fidelity/quality were associated with significantly *higher* (i.e. worse) disruptive behaviour scores at follow-up (ES = 0.37, CI = 0.02 to 0.72).
- Compared to low, and contrary to expectations, moderate levels of fidelity/quality were associated with significantly *lower* pro-social behaviour scores at follow-up (ES = -0.59, CI = -1.08 to -0.10).

**Dosage (H4c):** Findings pertaining to the relationship between levels of dosage and pupil-level outcomes were also mixed. Our analyses of the implementation data for the first year of the trial (2015/16) indicated that:

- Compared to low, high levels of dosage were associated with significantly *higher* reading scores at follow-up (ES = 0.27, CI = 0.02 to 0.52)
- Compared to low, and contrary to expectations, high levels of dosage were associated with significantly *lower* pro-social behaviour scores at follow-up (ES = -0.53, CI = -1.04 to -0.02)

Our analyses of the implementation data for the second year of the trial (2016/17) indicated that:

- Compared to low, and contrary to expectations, moderate levels of dosage were associated with significantly *higher* (i.e. worse) concentration problem scores at follow-up (ES = 0.74, CI = -0.14 to 1.60)

**Participant reach (H4d):** The association between reach and pupil-level outcomes was limited and negative. Our analyses of the implementation data for the first year of the trial (2015/16) indicated that:

- Compared to low, and contrary to expectations, moderate and high levels of reach were associated with significantly *higher* disruptive behaviour scores at follow-up (moderate ES = 0.70, CI = 0.32 to 1.07; high ES = 0.39, CI = 0.12 to 0.66)

Our analyses of the implementation data for the second year of the trial (2016/17) indicated that different levels of reach were not associated with any pupil-level outcomes.

**Participant responsiveness (H4e):** The association between levels of participant responsiveness and pupil-level outcomes was consistently positive, albeit limited to our analyses of implementation data for the second year of the trial (2016/17):

- Compared to low, moderate levels of participant responsiveness were associated with significantly *higher* reading scores at follow-up (ES = 0.24, CI = 0.02 to 0.46)

- Compared to low, moderate levels of participant responsiveness were associated with significantly *lower* concentration problem scores at follow-up (ES = -0.29, CI = -0.58 to -0.07)
- Compared to low, moderate and high levels of participant responsiveness were associated with significantly *lower* disruptive behaviour scores at follow-up (moderate ES = -0.34, CI = -0.63 to -0.05; high ES = -0.43, CI = -0.82 to -0.04)

Full models, along with accompanying sensitivity (Model 1.2) and MI analyses (Models 2.1 and 2.2), are provided in Appendix 12. In summary, our findings were sensitive to changes we made to our modelling parameters (e.g. modelling implementation data as continuous variables; MI of missing data), as follows:

- 2015/16
  - Modelling implementation data as continuous variables (Models 2.1 and 2.2) rendered the association between fidelity/quality and reading scores non-significant
  - MI of missing data (Models 1.2 and 2.2) rendered the association between fidelity/quality and pro-social behaviour scores non-significant
  - MI of missing data (Model 1.2) revealed a significant association between moderate (compared to low) participant responsiveness and higher pro-social behaviour scores at follow-up
- 2016/17
  - Modelling implementation data as continuous variables (Models 2.2 and 2.2) rendered the association between participant responsiveness and reading scores non-significant
  - Modelling implementation data as continuous variables (Models 2.2 and 2.2) rendered the association between participant responsiveness and concentration problem scores non-significant
  - MI of missing data while modelling implementation data as continuous variables (Model 2.2) rendered the association between participant responsiveness and disruptive behaviour scores non-significant

Given the relative sensitivity of these findings to changes in modelling parameters (at least, in comparison to those for H1, 2, 3 and 5) and the fact that several appear to be counterintuitive (e.g. higher levels of a given implementation dimension are significantly associated with worse outcomes), caution in the interpretation of analyses pertaining to H4 is advised.

### Qualitative case studies

The characteristics of the six case study schools are presented in Table 13. To note is that the overall case study sample broadly mirrored key trends observed in relation to the composition of the main trial sample (e.g. higher than average FSM eligibility, lower than average attainment). The diversity evident *within* the case study sample with regard to each school's characteristics is pleasing, especially considering that we did not use maximum variation sampling.

**Table 13: Case study school sample characteristics**

School	Area	Size	% FSM	% White British	% EAL	% Absence	% SEND	% Level 4+ English & Maths	OFSTED
1	Greater Manchester	Single form	Above average	Above average	Below average	Below average	Average	Average	Requires improvement
2	Greater Manchester	Triple form	Above average	Average	Below average	Average	Above average	Below average	Good
3	Greater Manchester	Single form	Above average	Below average	Above average	Above average	Average	Below average	Good
4	West Yorkshire	Single form	Average	Below average	Above average	Below average	Above average	Below average	Good
5	South Yorkshire	Single form	Above average	Above average	Below average	Above average	Below average	Below average	Good
6	Greater Manchester	Double form	Above average	Above average	Below average	Below average	Above average	Above average	Good

### Contextual profiles

School 1 is an urban Church of England primary school in Greater Manchester. It is smaller than average and is situated in one of the most deprived areas in England; the proportion of pupils eligible for FSM is twice the national average. The school is linked to another school in the local area, both of which are overseen by an Executive Head Teacher. Staff turnover between 2011 and 2015 was high, with almost all of the teachers being new to the school, including the assistant Head Teacher and four newly appointed governors. The school's most recent Ofsted grade of 'requires improvement' was based on the inspection criteria judgments about the 'quality of teaching' and the 'achievement of pupils'. In particular, Key Stage 2 attainment had fallen in comparison with previous years, and disadvantaged pupils did not make adequate progress. However, the school is currently exceeding the government's current floor standards regarding attainment; approximately three-quarters of pupils achieve Level 4 or above in English and Mathematics by the end of Year 6.

School 2 is situated in a deprived urban area and has a very large pupil intake. The school site itself is very small considering the number of pupils on roll, and staff have had to be inventive with the space available. For instance, break times are routinely staggered as there is not enough room for all pupils on the school's playground. Class sizes are somewhat larger than average because year groups are combined. For example, the Year 3 cohort participating in the GBG trial was combined with children in the year above in order to make three classes. The school has experienced a high turnover of teachers in its recent history, many of whom are NQTs: *"I am losing another member of staff this term, to be replaced possibly by another NQT, which would mean two now in the [Year] 3-4 team which is not where we started from"* (GBG co-ordinator). This high rate of staff turnover was a challenge that the school had had to deal with from the outset of implementation, with one teacher who had attended the initial training leaving halfway through the first term. This may have been one of the factors that led to the school ceasing implementation at the start of the second year of the trial.

School 3 is very ethnically diverse, and serves a deprived urban area with a high level of socio-economic challenges in the local community. Field notes from our first visit highlight the severity of deprivation: *"There is a lot of abandoned furniture and used needles strewn across the streets surrounding the school"*. The school has a very high FSM uptake. At the start of the first year of the trial, the school site was undergoing reconstruction, as there were long-term plans to expand its capacity to two-form entry. The building work appeared to be finished at the start of the second year, and did not seem to have

much impact on day to day running of the school. The school is part of an academy trust, and works closely with another primary school in the vicinity that was also part of the trial and randomly allocated to the intervention arm. As school 3 is part of a trust, an Executive Head Teacher has an oversight role, but the head of school conducts its day-to-day management. The schools in the trust work very closely together, exemplified by the sharing of school policies (including behaviour management) to the planning of lessons taught. School 3 does not have a high staff turnover, or a high number of NQTs. However, the Year 3 teacher was relatively new to the teaching profession, and had two higher-level learning support assistants (HLSAs) providing additional support in the classroom. In the second year of the trial, a more experienced teacher took the class, but the HLSAs remained to support individual children.

School 4 is an urban Roman Catholic primary school in Yorkshire. It is smaller than average and is notable for the ethnic diversity of its pupil intake. Only one-third of pupils are classified as being of white British ethnic origin, and almost half speak English as an additional language (with approximately 27 different languages being spoken in the school). Although the number of pupils with an SEN statement or EHC plan is low, the proportion of pupils receiving some form of support for SEND is much higher than in most schools. Additional provision was therefore provided by EAL and SEND teachers and teaching assistants. Prior to the trial, the school had experienced high staff and pupil turnover, with 45 staff starting and leaving the school between 2011 and 2014. In 2011, following an Ofsted inspection, the school was judged as 'requires improvement', and a new Head Teacher was subsequently appointed. Prior to this the school had been below national floor targets for seven years. More recent results showed approximately three-quarters of the pupils achieved level 4 or above in English and maths by the end of year 6. In its most recent Ofsted inspection the school was upgraded to the 'good' category.

School 5 is located in the middle of a council estate on the fringes of a large, deprived town. The school building is very small, and mobile classrooms had been erected in order to accommodate classes, despite the intake only being single form entry. The pupil population is predominantly White British, and speaks English as their first language. Reading scores are considered to be well below the national average, but writing and maths meet the national standards. There are strong ties with external agencies, the local secondary school, and the Sure Start centre that attached to the school. School 5 is part of an academy trust, with an Executive Head Teacher who has overall oversight and a Head Teacher who conducts its day-to-day management. The GBG co-ordinator was the Deputy Head Teacher of the school in addition to being head of Years 3 and 4; she left partway through the first year to take maternity leave. The Year 3 teacher had considerable teaching experience in the secondary sector, but was new to primary teaching, and was signed off on long-term sick leave at the end of the first year of the trial. The school therefore decided to cease implementation. The *"main reasons seem to be linked to staff turnover and therefore the school's Head doesn't feel comfortable to continue with it"* (school 5, email communication from Mentor).

School 6 is an urban Roman Catholic primary school in Greater Manchester. Almost all of the pupils speak English as their first language, with only a small proportion of pupils registered as English language learners. The school uses a streaming model, with higher and lower ability classes in each year group. It is situated in one of the most socio-economically deprived areas in England and the proportion of pupils eligible for FSM is therefore much higher than in most schools. Drug/alcohol addiction and domestic violence are critical issues affecting the local community. The majority of pupils live in local authority housing. As a means to address these issues, the school employs two learning mentors who work with parents and families, and a counsellor visits every two weeks to work with pupils. The school also runs breakfast, after-school, and holiday clubs, and is open 51 weeks of the year. Overall and persistent absence rates are higher than average, an issue highlighted in recent school inspections. However, attainment currently meets the government's floor standards, with two-thirds of pupils achieve Level 4 or above in English and Mathematics by the end of Year 6.

## Implementation

### Fidelity and adaptations

Consistent with the quantitative implementation data reported earlier in this chapter, teachers reported adhering to the procedures outlined in the GBG manual (e.g. reviewing rules prior to the commencement of the game, utilising voice levels, discussing exemplars of appropriate behaviours, providing praise to winning teams). Some minor procedural adaptations were made. For example, several teachers reported issues with the use of booklets to record game results for children, commenting that they “*hate the books*”, as they “*disrupt the flow and...it’s very confusing for the kids*” (school 3, teacher A) and “*take an awful lot of learning time up*” (school 3, teacher B), and so these were no longer used – an intentional, reactive adaptation based on their professional judgement.

Other intentional adaptations were more substantive, but were perceived by teachers as being necessary in order to meet the needs of their class. For example, some teachers interacted directly with pupils (or had a learning support assistant who sat with certain groups of pupils to support them) while the game was being played. One teacher explained that, “*I’ve had to intervene because I’ve got the lower ability class, so my children are SEN, so if they ever got stuck any point or kind of really didn’t understand what to do then I’d just I’d just intervene that way*” (school 6, teacher C). Another teacher commented that following advice from their coach, they gave a “*very small explanation*” to pupils following an infraction “*as to what they might be able to do or what they are doing wrong just to help*” (school 6, teacher D), as they felt that some lower ability children in their class were unaware of infractions they had made and so could not rectify the situation. Another discussed how they had adapted the pre- and post-game procedures for logistical reasons: “*condense it down, I don’t always go through the rules all the time because they know them inside and out...so we’ve probably chopped it down to suit us*” (school 1, teacher E).

Most teachers made adaptations to the reward system, although the types of adaptations varied, and these were broadly in line with GBG principles. Thus, while some teachers retained the use of stamps or stickers, others developed their own reward system based on the, “*need to move on to intrinsic rewards*” (school 8, teacher F):

*“I think initially I went for the kind of pencils and pencil toppers, but... it became an expectation then that behaviour equated to... monetary reward and I didn’t want the kids to get into the routine of expecting that kind of reward, so we’ve we worked a system at the minute that is a points based system where they’ve got to score thirty points per day to achieve a kind of expectation.”* (school 5, teacher G)

Other teachers developed different ways of rewarding their pupils; one recorded every time a team won a game, and provided a “*big reward*” after a certain number of wins, explaining that they wanted, “*to show more positive reinforcement...to give the children more ownership of the game*” (school 6, teacher I). If the pupils got “*a big prize every week then they’ll probably start thinking ‘oh I don’t care it’s only a pencil’...we want it to be an exciting thing for them*” (school 6, teacher D). Another school did something similar, whereby for each successful game, pupils received a marble that they could spend in the GBG “shop” on a Friday afternoon, otherwise “*they’d be going through pencils like crazy*”, and it teaches “*them a bit about saving up*” (school 1, teacher h). It is noteworthy that these adaptations are not incongruent with the expected evolution of the GBG in respect of delay of gratification and the shift from extrinsic to intrinsic motivation for positive behaviour in the classroom that would be indicative of generalisation of learning over time:

*“then when they get to the end that's when they choose the big prize as well.”* (school 6, teacher I)

The four GBG rules were adhered to, albeit with additional clauses/variations. For example, in some classrooms, rule three (‘we will get out of our seats with permission’) was varied to stipulate that, *“one person at a time from each team can be up [and] out of their seats”* (school 4, teacher J). Other variations included the use of *“TNT”* (tummies near table), *“BBC”* (bums and backs on the chair) and *“six legs on the floor”* to add specificity to the behavioural expectations underpinning the rule (school 3, pupil focus group). One teacher explained how they had initially been applying the rules *“in a very literal sense”* but was told by their GBG coach to be *“a bit more adaptable”* and so discussed the need to apply the rules, *“so that they make sense to the kids that you’ve got in front of you”* (school 5, teacher G).

Teachers also generally adhered to the prescribed procedures regarding the formation of teams, organising them *“to make sure it was a mixed ability kind of group”* (school 6, teacher D), and changing them around at various points in the school year. A variety of strategies were used to select team leaders. Some teachers selected *“higher ability”* children (school 2, teacher L) or those who *“would be the most responsible and able to go and get the booklets without a fuss and being the sensible ones to make sure everyone’s doing what they should be doing”* (school 6, teacher I). However, others deliberately chose children who were *“not necessarily usually the team leader that you would choose”* (school 6, teacher D), such as those *“who don’t excel academically...to give them more responsibility and a bit more, a bit of a confidence boost”*, or children who had, *“behaviour problems to try and encourage them to lead their team”* (school 4, teacher J). As with team membership, team leadership was periodically changed, *“to let them all have a go”* (school 1, teacher h).

## Dosage

### Duration

In line with the online scoreboard data, most teachers reported playing each game for between 10 and 15 minutes at the beginning of the year, although one suggested that if pupils *“can have half an hour on the Good Behaviour Game they tend to be settled”* (school 1, teacher h) for the rest of their lessons. By the end of the school year, teachers indicated that, in line with implementation guidance, they had *“been trying to play it for a longer period of time”* (school 6, teacher D), between 25 and 45 minutes<sup>15</sup>. The duration of game play depended largely *“on the activity itself”* (school 1, teacher E). For example, *“they need longer to do the writing... if I want them to do, for example, Big Write, I’ll give them forty minutes where they’re working in silence and so it just depends on the activities”* (school 6, teacher I). Another teacher explained, *“it varies from like ten minutes if it’s something like we did today, like drama, or something a bit more active then it tends to be shorter, up to... forty minutes if it’s a writing task or something like that”* (school 4, teacher J).

Game duration was reportedly shorter in the second year of the trial, with one teacher explaining that even by the end of the year, they still would only *“tend to do anything between ten and fifteen minutes. I find sometimes if you go more than fifteen minutes that's when they start losing it a bit”* (school 1, teacher E). Another explained that to meet the needs of their lower-ability class they, *“have done shorter ones and...have done a couple longer ones but fifteen, twenty minutes are normally the best kind of time that I’d set an activity for the kids, you know what I mean”* (school 6, teacher C). Thus, the average

---

<sup>15</sup> While this is highly incongruent with the overall trends observed from the online scoreboard data, it is possible that teachers in case study schools were among the small number playing for longer than 15 minutes.

fifteen minute 'ceiling' on game time observed in the online scoreboard data was set based on teachers' expectations and judgements regarding the abilities and needs of their classes.

### Frequency

At the outset of implementation, teachers stated that they were playing the GBG with their pupils at least three times a week, although some reported problems in, *"making sure that it's integrated consistently"* (school 5, teacher G) due to competing priorities. As the school year progressed, teachers reported that they *"play it for more often"* (school 4, teacher J), with most aiming for between three and five times a week in both years of the trial. However, this varied considerably. Thus, by the end of the year, while some teachers stated that the GBG was *"happening on a regular basis, happening every day"* (school 1, teacher E), others noted their game play had reduced to, *"maybe twice a week...just because of other things that are happening within the school...it's other things within the school that I'm finding really hard to fit in those games, especially the longer ones"* (school 6, k). There were also some disparities between teacher and pupil reports regarding the frequency of implementation, with the former generally reporting higher frequency than the latter.

### Timing and subjects

There was no clear consensus over the best time to play the GBG, with teachers citing different preferences regarding lessons and times in the day. Generally, they chose lessons in which they felt their pupils were strongest in order to optimise their chances of success, and/or in which they felt the game procedures aligned well with the nature of activities undertaken. One teacher said they had *"started using it within my English lessons"*, but found it *"more difficult in other subjects, for instance, Maths, because sometimes you need the children to talk to each other"* (school 2, teacher O). Conversely, another suggested that while the GBG *"works really well in Maths because we can use it for a lot of starter activities, ... it doesn't work as well in English because they need a lot of support... they're not as strong at English as they are in Maths"* (school 1, teacher H). Others expressed a preference for playing the GBG *"for English and Maths because they are more structured lessons"* (school 6, teacher C). As the year progressed, teachers cited use of the game throughout the day, in a range of lessons. A couple had also taken the game out of the classroom, *"in the hall for when we've done PE and for assemblies"* (school 1, teacher E) and one had, *"tried one music lesson in the hall"* (school 6, teacher C).

Decisions about the timing of game play were also the result of its perceived calming effect on pupils. It was seen as ideal for transition points in the day (e.g. at the beginning of the day, after lunch) as it *"kind of calms them down"* (school 2, teacher L) and could therefore be *"a really constructive way to get them refocused in the afternoon and ready"* (school 5, teacher G). As the year progressed, some teachers chose to play the game during tasks that would require their pupils to move around or talk, in order to challenge them further and develop their teamwork skills.

### Participant Responsiveness

Teachers across the six schools commented on their pupils' affection for the GBG, with some reporting that they frequently requested, *"can we play the Good Behaviour Game?"* (school 2, teacher O) and would, *"get so excited about playing it"* (school 3, teacher A). However, one teacher felt that during the second year of implementation, the pupils were, *"a bit bored of saying the rules [laughs]. I think they find that a bit sort of repetitive"* (school 4, teacher J). Consistent with teachers' accounts, pupils themselves generally reported considerable enjoyment of the GBG, expressing that it was *"the best game, learning game, in the world"* (school 4, class of teacher J, pupil focus group), as it *"means you get to do more fun things"* (school 6, class of teacher D, pupil focus group). Specifically, they *"like it when you work in partners"* (school 3, class of teacher A, pupil focus group) and *"like doing the celebrations"* (school 1, class of teacher E, pupil focus group). However, some also commented that

although it was “*easy once you get used to it*” (school 1, class of teacher E, pupil focus group), they found the GBG to be initially rather challenging, particularly in games involving the use of ‘voice level 0’ for activities in which, “*you really, really need help*” (school 4, class of teacher J, pupil focus group). Finally, there was some evidence of habituation, in which the game became “*boring*” (school 1, class of teacher E, pupil focus group).

There was also evidence of perceived differential responsiveness, although there was no clear pattern as to which groups of pupils were viewed as being more or less engaged with the GBG. For example, while more academically able children, “*get a little frustrated at the beginning when they’re waiting to get started with their work*” (school 4, teacher J), equally, lower ability children “*can be quite disengaged...if it is a big task like assessed writing...they do struggle*” (school 3, teacher A). Another teacher explained that the GBG allowed the lower ability children to “*coast*” as they could, “*switch off a little bit and... give up the responsibility, hoping that the other team members will pick it up*” (school 6, teacher C). Conversely, one teacher felt that children at all levels of ability, “*responded in a similar way*” (school sa776, teacher H). Losing the game was reportedly an issue for some pupils, particularly if they are “*quite emotional*” children (school 3, teacher A). Despite this, teachers found that even if there was an initial negative response to losing, pupils would quickly “*move on*” (school 4, teacher J) and can be “*very mature about it*” (school 6, teacher D).

#### *Programme Differentiation*

Consistent with the usual practice survey analysis reported earlier in this chapter, teachers saw considerable similarity between their existing behaviour management strategies and the principles of the GBG, most notably in relation to the focus on rewarding positive behaviour, and the emphasis on classroom rules. Most felt that the four GBG rules were so closely aligned with existing classroom rules that, “*it is not an issue at all*” to combine them (school 4, teacher J). Even in situations where teachers chose not to formally merge the GBG with their existing systems, “*they seem to be sort of mutually supportive of one another...our behaviour chart is reinforced by the Good Behaviour Game...so the two are quite ... they’re well matched*” (school 2, GBG co-ordinator). Only one school felt that the ethos of the GBG did not align with their existing behaviour management strategy, which was more sanction-based. However, the teacher felt that the use of four infractions in the GBG meant that, “*children can make a mistake and learn from it*” (school 1, teacher H).

Teachers tended not to make evaluative comparisons, although a couple commented that the GBG was “*so much better*” than their previous strategies (school 6, teacher D), and emphasised the usefulness of the voice levels when trying to settle children down, remarking that it was “*so much easier*” (school 2, teacher L) than other strategies that they had used previously. One reported that they had called upon on their previous practices as a trainee teacher to create a positive reward system as an adaptation to the GBG. In sum then, the level of programme differentiation was deemed somewhat low, with mixed implications. On the one hand, familiarity and experience with key principles of the GBG provided the footing for a smooth transition to implementation. However, from another perspective, this may also have limited the achieved relative strength of the GBG and thus, the likelihood of measurable impact on academic and behavioural outcomes.

#### *Programme Reach*

The GBG was implemented “*for all the groups, all the children in the class*” (school 6, teacher C) for the most part. However, in certain circumstances it was used as an opportunity for withdrawal work with some children: “*we have an in-school counselling session and we have an EAL teacher come in, if they’re requested to go to that appointment then I have to let them go so those children would be absent*” (school 3, teacher B). However, importantly, pupils were not explicitly removed for reasons relating to their ability to engage with the game (e.g. those with significant behavioural difficulties). Instead, the



support of learning support assistants was utilised, particularly if it is a “big task” or a “really long” game, as a means to prevent them becoming “disengaged” (school 3, teacher A).

### Quality

Our interview data revealed a range of approaches taken in order to enhance the delivery of the GBG. Prior to implementation, several teachers prepared their pupils for the game by utilising strategies such as, “Circle Time around the rules and expectations” (school 5, teacher G), and “made it sound quite exciting... made them as involved as possible in terms of choosing what their rewards are, and just linked it in with...class rules and school rules” (school 4, teacher J). At the start of implementation, increased emphasis was placed on the pre-game procedures in order to ensure pupils’ understanding of what was expected of them:

*“We’d go through each rule and we’d say, at the start I spent a lot longer and I don’t need to spend that time now, so at the start I’d spend a lot of time like, and you know, on the voice levels...what I expect, we spent a lot longer going through all those different things and talking about good examples and not and showing that, whereas now they know what’s expected.”* (school 6, teacher I)

Similarly, end of game ‘debriefs’ were held to discuss with pupils what had not gone well, and what could be done to improve on this. For example, one teacher recounted how they had asked their pupils to work at a higher voice level to “help each other all the way through”, but they had “just focused on themselves... [so] after that we had loads of discussions about why we use that voice level and, how we could possibly help each other” (school 6, teacher D).

Teachers also actively involved pupils in decisions regarding implementation, allowing them to choose preferred rewards, and “discuss it with the children about what we’re going to do next with the Good Behaviour Game” (school 4, teacher J). Games were incorporated into lesson planning, with teachers looking for opportunities where the game aligned well with the requirements of a lesson (e.g. when a task required silence or collaborative work). Some used the GBG as an opportunity to teach foundational skills for life and work such as “saving up [money]” through token-based reward systems, “the value of really working together”, and “helping each other out and taking responsibility” (school 1, teacher H). Marginal adaptations to accommodate pupils with additional needs were also evident:

*“Not that we’re necessarily more lenient with them, but obviously I know that if they were doing something I wouldn’t necessarily count it as an infraction as I might do possibly for somebody else. I’ve got one boy in class who can make noises and can be quite loud but that’s just the way he is that’s his nature so if it was somebody else that would be a rule break but... that’s who he is so... in some ways you make some allowances for those with additional needs.”* (school 1, teacher E)

## Factors affecting implementation

### Intervention characteristics

#### Voice levels and rules

The extent to which key principles and procedures in the GBG aligned with teachers’ beliefs, values and preferences regarding behaviour management were central to their willingness to implement it as planned. In both years of the trial, the majority of teachers felt that the ‘voice levels’ concept was useful, and integrated it into their usual practice “when the game’s not being played from saying like ‘use level zero’...[it] is working a lot more effective than just saying like ‘be quiet don’t talk’” (school 6, teacher I).

However, a minority did not find it useful, stating that they expected their pupils to already know what an appropriate noise level would be for a given task: *"We don't use [it] an awful lot, we just expect this...you know, this level, get on with it... we never said like 'Oh well this is where you whisper...' we just expect probably the same all the time"* (school 6, Head Teacher).

The four GBG rules<sup>16</sup> were also well received, primarily because of their simplicity, generalizability and applicability (*"it's really good because like 'following directions' is just everything that you do...and then the 'being polite to others' as well that's such a nice one"*, school 6, teacher D; *"[by saying] 'is that being polite'?...and it reminds them straight away about the Good Behaviour Game"*, school 6, teacher D) and alignment with existing practice (see 'programme differentiation' above). This led to their integration beyond the game: *"[we] use the four rules as model vocabulary now"* (school 3, teacher B). The positive framing of the rules was also seen as helpful: *"the rules are positive as well as opposed to 'you're not allowed to do this, you're not allowed to do that' and it works really well with the kids"* (school 3, teacher A).

#### *(Lack of) direct interaction*

However, as noted earlier (see 'fidelity and adaptations') most teachers in both years of the trial found the mandated lack of direct interaction with their pupils during the game challenging, as *"it's hard to not have an input as much during the day where I've just got to walk round...I find that a bit tricky"* (school 1, teacher H). At the start of implementation, particularly during initial training, this was perceived by teachers as being a particularly problematic element of the GBG: *"I was thinking I'm not sure how this is going to work because when I saw the videos there was just a teacher walking around the classroom and it was complete silence...I was thinking 'you can't do that in the middle of Maths'"* (school 1, teacher H). They expressed concern about the implications for their ability to support and clarify understanding with pupils that struggled with the task at hand:

*"Initially I played the game rigidly but it became demotivating for the really low achieving students who, as much as I wanted to encourage to work independently, and that's exactly what I wanted to do, some just couldn't access it. I mean we've had a boy that's joined recently who's not accessing Year 1 you know, so it's difficult enough differentiating down to... [that] level."* (school 5, teacher G)

However, over time, some of these concerns waned as teachers became more comfortable with the increased autonomy children experienced during GBG, and *"now I can pop it into lessons, still struggle a little bit with certain lessons and certain activities but it is it's affecting us less and less as we go on"* (school 1, teacher H).

#### *Data monitoring procedures*

The data monitoring requirements of the GBG divided opinion. A minority found the online scoreboard problematic: *"finding the computerised version difficult... [because it involves] turning your back on the class"* (school 6, teacher D). However, others greatly preferred it to filling in paper logs and *"that's a lot better now it's on the whiteboard"* (school 6, teacher I) as *"the paper one... was quite time-consuming filling in... making sure it was photocopied for when [coach] came in"* (school 6, teacher I). The online scoreboard was viewed as benefitting pupils too, *"because it gives the children a visual"* (school 3, teacher B).

---

<sup>16</sup> (1) We will work quietly; (2) We will be polite to others; (3) We will get out of our seats with permission; (4) We will follow directions.

However, teachers across both years of the trial had strong views about the GBG booklets used to record game results for individual children, going so far as to say *“I hate the books”* (school 3, teacher A). The lengthy process of stamping individual books was deemed to be so time consuming that it inadvertently defeated the purpose of the game: *“they were just taking up too much time and I found like even just that whole transition from the game itself like back into regular classroom used to get really noisy”* (school 3, teacher A). The layout of the booklets was not viewed as intuitive for pupils: *“it’s very confusing for the kids - they just stamp in random places”* (school 3, teacher A). This led to teachers either omitting the use of booklets or using alternative strategies to save time, *“I just say ‘team leaders put the stickers on jumpers’ and the children are still getting the reward, they still get a sticker”* (school 6, teacher D).

### Rewards

The reward system that is so central to the GBG was received favourably by teachers, who were routinely using similar systems in their usual practice (see ‘usual practice’ above). However, they felt it was important to involve children in making choices about which rewards to use: *“giving them the ownership to choose the prize as well, so it’s so much more than just a sticker”* (school 6, teacher I). Many used this method to facilitate the transition between tangible and intangible rewards, as well as increasing the delay in gratification by providing pupils with opportunities that they would not normally have: *“the ultimate goal is my spinny chair...it’s the leather padded spinny chair...so they can buy that for a lesson...we’re trying to get them away from pencils to things like that and then we’re going to be doing like more intangible stuff”* (school 1, teacher H).

## Programme support system

### Training

Although the initial GBG training was deemed to be useful, many found that it focused too heavily on the theoretical principles: *“it would have been useful from our point of view... to explain the game and its principles before they went into everything else afterwards”* (school 6, Head Teacher). Similarly, *“it would have been nice to see the game... at the start... I was a bit confused what it was until the second day”* (school 6, teacher I). This led to a perception that two full days of training was not necessary: *“it could have been quite easily reduced into one day”* (school 5, teacher G). A streamlined version, focusing primarily on the practicalities of implementation, would have been preferred, because *“as soon as they started demonstrating how to do it...then you got practical questions and...it was more relevant”* (school 4, teacher J). Thus, seeing and experiencing the game was viewed as fundamental in terms of preparedness to implement: *“the fact that we could see the game being played, we could play it on the day made it easier to visualise it...with regards to using it in the classroom...it would have been different [if] we’d just talked about it all day...and not actually played the game I don’t think I’d have come away and been able to implement it quite as well as I have done”* (school 2, teacher M); *“[watching] the videos and actually being able to see it being played...it was really easy to then model it in the classroom”* (school 6, teacher D).

Follow-up training was welcomed by schools as, *“it was nice to hear the other stories from other schools...about what was working, what wasn’t working”* but they did not feel that they *“learnt...anything deeper about the Good Behaviour Game”* (school 6, teacher D). Thus, it was primarily beneficial because it enabled schools to share experiences, challenges, and good practice: *“it was useful to find out what other schools were thinking of it...and getting ideas from them”* (school 6, teacher I), and, *“it was good to speak to other teachers because I was thinking...maybe it’s just me having certain struggles...but it was good to get to speak to other people who found the same thing [difficult]”* (school 1, teacher H). This contributed to a renewal of motivation: *“it was kind of nice to feel enthusiastic...like we did at the beginning again so that was good”* (school 6, teacher D).

### Coaching

Teachers in the first year of the trial expressed initial scepticism about the coaching aspect of the GBG, but this faded over time: *“honestly at first I was like a bit like ‘oh somebody else to come and watch me’ ...that’s not been the case at all it’s just been someone who’s wanted to see the game and help”* (school 6, teacher I). Regular visits from the school’s GBG coach had a reassuring effect, as *“I find that when I spoke to her it boost your confidence in like ‘oh you are doing it right”* (school 6, teacher I). This aided the development on a robust and transparent working relationship: *“I know that I can ask her and there’s no come back so to speak, that she’s not going to judge me on, you know, if something goes slightly wrong”* (school 2, teacher O). Coaching conversations also served the purpose of reinforcing the importance of a consistent approach to implementation because, *“that’s the reality in schools isn’t it, things do get forgotten and priorities take over and I think what a coach does is keeps it on track and keeps it going”* (school 5, Deputy Head Teacher). The coaching conversations were also a chance for the teachers to consider how they could develop their GBG practice and create new challenges for their class, *“she’s showed me all those different ways of doing it”* (school 2, teacher L); and *“...she’s always wanting you to take it like a bit further”* (school 6, teacher I). Thus, by the second year of the trial, there was no scepticism evident, and the intended function of the GBG coaches was fully understood:

*“[The coach is] really positive with the feedback that she gives, she sends quite detailed reports back through...she always offers you that that next step with where to go with it next.”* (school 3, teacher B)

The fact that support extended beyond coaching conversations and was offered flexibly around teacher’s needs was similarly appreciated: *“she’s always available if we want any information...I asked for a meeting with her and she was here within a day so it’s, you know, it’s been very helpful to have somebody there”* (school 2, teacher L).

### Classroom level factors

#### *Pupil needs and attitudes*

A key challenge identified in relation to the needs of particular children in participating classes was their relative lack of autonomy. The GBG was seen by some as a means to address this, as *“they’re very needy and not independent at all so we thought this was perfect for them.”* (school 5, Deputy Head Teacher). However, as noted earlier, this also created tension in relation to teachers’ natural inclination to provide direct support during tasks. Without this extra support, it was felt that some pupils simply could not access and engage with the GBG, *“there’s a lot of children in my class that can’t sit down and do that on their own...without a teacher no matter how much input I give to them before the Good Behaviour Game”* (school 6, teacher D). Focusing on the teamwork element of the intervention as a means to strengthen collaboration between children was viewed as one way to compensate for this:

*“I was trying to challenge the children to use teamwork...the children found it hard to meet but if they did win then they seem to learn their lesson for the next game”* (school 6, teacher D).

The aforementioned teamwork aspect of the GBG was also favoured by pupils, who felt that it created a friendly dynamic when working (*“we always work as a team and the way everyone’s really nice”*, school 6, class of teacher I, pupil focus group), and were very aware that the GBG had reinforced their pro-social skills and the importance of being a good team member (*“it’s good because it helps you work as a team and it helps you be more friendly”*, school 6, class of teacher I, pupil focus group). Pupils’ understanding of the importance of teamwork in order to “win” the GBG (*“you should be ashamed of yourself if you break a strike and you let your down your team”*, school 4, class of teacher J, pupil focus

group) and their enjoyment of the intervention were contributory factors that made implementation easier for the teachers: *“The children absolutely love it so it’s always a bonus when it’s something the children like doing”* (school 2, teacher M).

As would be expected, rewards were strong motivators (*“You want to not get five marks because you really want to get that prize”* (school 4, class of teacher J, pupil focus group). This was routinely identified as pupils’ favourite aspect of the game, *“it’s really fun ‘cos you get surprises and all that”* (school 5, class of teacher G, pupil focus group), and the importance of the social aspect of the process was recognised by their teachers: *“they respond to praise as well like they like to celebrate it publicly in the class as well as getting their counter”* (school 3, teacher A). This was in part because, *“they’re all very competitive”* (school 2, teacher L).

However, working in silence – voice level zero – was seen as a particular challenge. Pupils reasoned that when struggling, they couldn’t ask others for help as this would lead to an infraction: *“It’s really hard because the times when we don’t get what we’re doing and we’re on voice level zero we try and tell our friends but they can’t either and we want to tell our teacher but we can’t.”* (school 6, class of teacher I, pupil focus group). In such circumstances, a common strategy was to just attempt the work, thus risking failure (*“you just have to try and try it”*, school 6, class of teacher D, pupil focus group). Coupled with the mandated lack of direct intervention on behalf of the teacher, this led to difficulties for some:

*“I was stood behind her saying ‘team number two you’ve broken rule number four not following direction, everyone else well done for following direction’ and the rest of her team just sat and didn’t even look at her...I literally said that about six times in a row and I felt so sorry for the girl because I couldn’t actually say ... ‘come on the rest of your team you need to help her’.”* (school 6, teacher D)

#### *Teacher attitudes*

Teacher interviews revealed a consistent pattern of initial scepticism about the GBG that appeared to be rooted in a preference for existing practice: *“the honest opinion was...I know what my behaviour management is...I thought I don’t want to be using something else which I’m unsure of because I know they I like my children to behave”* (school 2, teacher L); and *“I looked at the paperwork and I’ve got to admit I didn’t like it at all”* (school 6, teacher D). However, once implementation began, their perceptions shifted as they realised that assumptions they had made were incorrect:

*“It’s been a great tool to use as an extra behaviour management strategy and like I say I want to implement it more into my teaching...when you actually read about it’s different to [your] first perception of it, you’re like ‘oh I don’t want to do the Good Behaviour Game’ but it does work.”* (school 2, teacher L)

*“Once you get your head around it, easy...The thing for me was to understand that to make it part of what you do already so once I got my head around that I was fine because it’s not an extra thing it’s just...something you do as well.”* (school 6, teacher I).

Another attitudinal factor that affected teachers’ implementation of GBG throughout the trial focused around the importance of choosing meaningful activities set for GBG. Teachers admitted to feeling a certain pressure to select certain activities in order to meet dosage requirements, *“...just occasionally...you kind of feel like you need to shoe horn it in somewhere...and it’s really tempting to just do one for the sake of doing one”* (school 4, teacher J). However, teachers felt very strongly that GBG had to be planned around activities that challenge their pupils and would rather lower the frequency of implementation if lesson plans did not align with the intervention, *“I could always fit it in but I don’t want to fit it in where it’s not just an easy win for them...I don’t want to just do [it] during a spelling test you know...so that’s why I don’t always do it because I don’t want it to just be...another*

tick" (school 4, teacher J). Teachers felt this approach worked as GBG optimised the selected activity for the children so they would get the best possible learning opportunity, "*sometimes it's peace of mind as well if you think 'I'm going to do The Good Behaviour Game for that activity' you know that the children are going to try so hard...and do a really good job so...I really enjoy it*" (school 6, teacher D).

Staff also commented on the goodness of fit between the GBG and the need for structure among certain groups of children:

*"You can't teach a classroom without structure within it and a lot of our children if that structure isn't there will seek to misbehave, not necessarily because they're disengaged at that point but...they need to know...what is expected of them. They relax when they know what's expected of them if the lessons are pitched at the right pace they engage and they make progress. I think what the Good Behaviour Game does is help reinforce that in a different way, it's not just the teacher's rules it's a game so it reinforces our behaviour management system but by making it fun at the same time"* (school 2, GBG co-ordinator).

*"Our school has a certain level of.... deprivation... some of those factors... can bring themselves into the school so we thought it would be a good idea to give them some structure to help them modify their own behaviours and that [the GBG] seemed to fit that sort of ethos."* (school 2, GBG co-ordinator)

However, in contrast to the above, one teacher stated, "*we've got so much SEN...I think personally it would be unfair to penalise really low achieving pupils for not being able to interact, especially with a teaching assistant*" (school 5, teacher G). Thus, the mandated lack of direct interaction during gameplay was noted once more, this time presented as a barrier to meeting the needs of struggling pupils:

*"I think Numeracy's the one I find most difficult because my we have two TAs in the classroom and very low children are just used to an adult being with them all the time and encouraging them or boosting their confidence...or even just kind of modelling through what they need to do and I find it really difficult to do an independent Numeracy task."* (school 6, teacher D)

### *Planning to play*

Perhaps unsurprisingly, teachers were less likely to implement the GBG during busy periods in the school year, in which there were scheduled school events or assessments: "*next week we've got lots of Christmas stuff... going on so we might not play it as much*" (school 2, teacher M); and, "*when it comes into like assessment weeks it can sometimes get shelved because you've got priorities that you've got to hit*" (school 5, teacher G). In such circumstances, the role of the coach in holding teachers accountable for consistent implementation was seen as critical by some: "*coming up to assessments and stuff, sometimes it's a little bit easier to push to one side...so having your mentor coming in again and again kind of...brings it to [the] front a bit more*" (school 6, teacher D).

At the start of the first year of implementation, many teachers felt that they would benefit from extra planning time in order to deliver the GBG well: "*if I had extra planning time I would probably play the game better...there's not necessarily all the time to sit...and think 'right I did that last time I'm going to do this'*" (school 6, teacher D). However, over the course of the year a shift took place, at the end of which the intervention was seen as more integral to existing planning: "*it seems to have kind of fitted...into the planning that we're doing already...I don't think we need extra planning time for it*" (school 6, teacher D). This continued into the second year of the trial: "*I don't think I need extra time to plan for the Good Behaviour Game because a lot of the games that we do play are part of our normal teaching anyway*" (school 1, teacher N). Thus, as teachers became more familiar and comfortable with implementing the GBG, the amount of effortful planning time was reduced. There was a shift from fitting

activities around the GBG, to fitting the GBG around activities: *“at first I was like ‘oh I’ve got to plan something different we’re playing the Good Behaviour Game’ whereas now I’m like ‘oh we’re doing such a thing oh that’s a good time to play the Good Behaviour Game”* (school 6, teacher I).

### School level factors

#### *Senior leadership team support*

GBG co-ordinators (who were also typically members of the SLT) were more prominent in the first year of the trial, adopting an oversight and monitoring role, *“I’ve seen one game from each class initially once they’d start running with it and it worked really well...I’ve popped my head informally when they’ve been playing”* (school 6, Head Teacher). Teachers felt that this form of support was adequate, but saw themselves as leading the intervention, *“they’ve come to see a couple of games but they’re not involved in the teaching of it...they do monitor it and check on how it’s going”* (school 1, teacher H). As the school year progressed, GBG coordinator support was slowly withdrawn (*“initially I had the Deputy Head who was involved, but I think she’s taken a bit of a back step”*, school 5, teacher G) and was almost entirely absent in the second year of implementation, with teachers not able to provide any indication of their role in interviews.

In both years of the trial, the support offered from other members of SLT was informal, as *“they just ask me how it’s going and if I need anything”* (school 4, teacher J). Staff felt they did not need direct support, but were comforted that they knew it could be accessed as required: *“to be honest it’s all got off on like a smooth start...but if we do need anything then we’ve got [Deputy Head] to talk to about things”* (school 6, teacher I). The most common form of interaction between SLT and implementing teachers revolved around updates on the implementation, so *“there’s an on-going dialogue in terms of impact and obviously they they’ve met...the regional specialist several times whose been in...it’s always been a collaborative thing...so yeah I feel supported in it by the school”* (school 5, teacher G), and *“the Head Teacher is involved...she’s gone to the training and she likes to have regular updates on what we’re doing”* (school 1, teacher H).

#### *School climate and openness to change*

Schools that felt their existing ethos and practices matched with GBG were more receptive to adopting the intervention and *“thought it would be a good idea while trying to raise attainment and progress within the children to give them some structures within that to help them modify their own behaviours and [GBG] seemed to fit that sort of ethos”* (school 2, GBG co-ordinator). SLT members in case studies schools were responsible for driving the vision for the GBG and hopes for integration across the school in the future. They expressed the importance of being receptive to new strategies in order to better meet the needs of their pupils:

*“We just thought it was a brilliant opportunity to see another strategy or another way that it could be used so that maybe we could implement it, ourselves afterwards or completely change our behaviour strategies...[the school is] quite open to anything really...[it] just sounded like the kind of thing that we are into when we initially read about it.”* (school 5, Deputy Head).

There was variability evident in the extent to which the teachers who would ultimately be implementing the GBG were involved in their school’s decision to participate in the trial. So, while some were actively consulted, *“it was our Executive Head Teacher, the Partnership basically came to us and asked us would we be willing to participate and we said we’d really like to, sounded interesting and we’d just like to see what it’s all about really”* (school 3, teacher A), others were simply informed of the decision, *“I was just told ‘you’re going on some training’ and off I went”* (school 2, teacher M). However, there was no evidence that this impacted on their attitudes toward implementation, with most expressing openness

to change, feeling that the GBG, *“sounded interesting and we’d just like to see what it’s all about really”* (school 3, teacher A).

### Perceptions of impact

Teachers reported observing differences in pupils’ learning and attainment, behaviour and social skills that they attributed directly to their participation in the GBG. The mechanisms underpinning these perceived impacts were consistent with the underpinning theories and logic model of the intervention outlined in the introductory chapter of this report. Thus, in terms of learning and attainment, they reported increased on-task behaviour (*“[they can] concentrate better”* (school 2, teacher O), autonomy (*“they have become a lot more independent”*, school 6, teacher K; *“their hands used to be constantly up, literally they’d write out their date and they’d say ‘I don’t know what to do’ but they know now they need to try and figure it out, they need to ask somebody beside them”*, school 3, teacher A), self-efficacy (*“in terms of changing attitudes to learning it has had significant impact... [they are] willing to risk now...if you can get a child to risk making mistakes then you’ve got a confident learner”*, GBG coordinator, school 4) and confidence (*“in terms of getting children to speak who don’t normally speak or wouldn’t put their hand up, if I make them the team leader then they have to”*, school 6, c; *“the quiet ones... come out of their shells a lot speaking and listening wise”* (school 6, teacher D). These improvements in adaptive learning behaviours were theorised to underpin changes in attainment, such that pupils had, *“made so much progress, even academic wise”* (school 6, teacher D), and that *“it improves their work... the quality of their work during the Good Behaviour Game I’ve found is better”* (school 4, teacher J).

The teamwork aspect of the GBG was seen as having impacted positively on pupils’ social skills and relationships. For example, one teacher commented that they had to provide *“less support in the way that smaller things that you would normally have to be around, like sorting out squabbles about rubbers and pencils, all that’s gone”* (school 1, teacher H), while another noted, *“you can hear more positive language from them so they’re like ‘oh that’s a good idea’, there’s no fighting or anything like that so I feel like it’s really, really working”* (school 6, teacher I). There was a sense in which the GBG had led to a more pro-social spirit, with pupils being more willing to *“help each other”* more (school 4, teacher J) and be *“more cooperative with one another”* (school 1, teacher E). Pupils were seen as increasingly using each other to support their learning as opposed to relying on their teacher: *“they know now they need to try and figure it out, they need to ask somebody beside them”* (school 3, teacher A).

The above impressions were confirmed by pupils, who perceived benefits in similar domains including on-task behaviour (*“it helps me concentrate when like when the noise level is set to like one I can concentrate”*, 1, class of teacher E, pupil focus group), social skills and relationships (*“it makes people learn how to work as a team”*; school 4, class of teacher J, pupil focus group; *“it helps you be more friendly”* (school 6, class of teacher C, pupil focus group) and in addition, improved self-regulation (*“helps us to be calm more often”*, school 6, class of teacher C, pupil focus group), with comparable consequent benefits in terms of learning and attainment (*“helps us to understand loads of stuff and make our writing better”*, school 6, class of teacher C, pupil focus group).

However, in spite of the above, there were differences of opinion regarding the extent to which the effects of the game were generalised. So, while some reported that, *“it’s basically developing your class’ skills when you’re not playing the game”* (school 4, teacher J), others noted that, *“as soon as the Good Behaviour Game finishes they seem to slack off a little bit”* (school 6, teacher K). In some cases, it was felt that pupils needed to be actively encouraged to generalise the behaviours that they learned during the game. Explicit promotion of adherence to key GBG principles, such as the four rules and voice levels, was central to this:

*“We try as much as we can to remind them of the four rules throughout all school all school day and all school life without it just being when it’s in the game... as much as possible we say, you know, ‘if we were playing the Good Behaviour Game that would*



*have been a rule break' and, you know, 'what does that look like when we're thinking about this rule' so we do really try and apply it across, you know, all of school life so yeah think and they are I think can really see the difference with that they are getting much better at that definitely."* (school 1, teacher E)

*"It works and it's kind of...spreading out to other areas as well its helping them because I can use those rules all the time I don't have don't have to just use them during the game"* (school 2, teacher M)

*"We do use the different voice levels now, we use [them] all the time throughout the day"* (school 5, teacher G).

However, these positive perceptions of impact were not universal, particularly among schools that ceased implementation. For instance, one teacher noted that when they *"did play it the behaviour was worse at the end of the game than it would have been if he hadn't started it, so [we] didn't really see the benefits"* (school 7, GBG co-ordinator). Another commented:

*"I did not feel that their work was better. I actually think that the work they produced was worse during the game. I found that basically hindered my children's learning opportunities and learning environment really. I wasn't happy with it and I feel like the work they produced was worse off because everything you had to give them had to be something that they'd done before. You cannot give them something that's new because you cannot help them but as a teacher that's what I'm there for."* (school 8, teacher F)

Furthermore, teachers disagreed with one another regarding whether perceived impacts were sustained. So, while one teacher stated, *"I think even without the teachers doing it next year I think they'll automatically, if you mention the rules to them, they'll automatically fall into, you know, what they know"* (school 6, teacher C), others felt that the pupils had not really improved from the previous year of exposure to the intervention, *"I guess some of the children, their behaviour, just with them getting older they're kind of questioning the boundaries a little bit more than they were in Year Three"* (school 4, teacher J).

### **Differential gains**

The inconsistency evident with regard to teachers' views on the suitability of the GBG for different groups of pupils and their levels of engagement with it (see, for example, 'Teacher attitudes' and 'Participant responsiveness' above) was also highlighted to a degree in their discussions of whether any of these groups benefitted differentially in terms of the outcomes noted above. Most felt that the game was particularly beneficial to their pupils with additional needs. For example, one teacher commented:

*"The child that I see the most benefit is the little girl with autism... the structure really helps them, so I've got I think I've got eleven children on the SEN register in my class but... because the rules are so explicit and because it like the directions are so explicit, they just really thrive under it...the children that it's been more beneficial to are the children who struggle when they don't have structure... like the children who are quite clearly SEN in my class... I think it's probably been better for them."* (school 6, teacher D)

Other teachers also used words such as *"thriving"* to describe pupils who they had previously had difficulties with, and a GBG co-ordinator believed:

*“It gives them ability to show different aspects of themselves that are in there with them but don’t always get the opportunity to let that come out, so while they might not be in the top group for Maths or something they can be in the top team that’s winning the Good Behaviour Game and that in themselves gives them self-esteem, which then enhances their learning overall.” (school 2, GBG co-ordinator)*

The structure and explicit nature of the game was also felt to be beneficial for lower ability groups or those who spoke EAL as *“they’re a lot clearer about what’s expected of them”* and they, *“take ownership...because it’s their responsibility if they don’t understand what they’re doing they’ll get strikes for the team so they’ve kind of switched on a little bit more and listening more”* (school 4, teacher J). However, other teachers disagreed, reporting that lower ability pupils would *“switch off a little bit”* (school 6, teacher C), and would particularly struggle with the longer games:

*“There’s a lot of children in my class that can’t sit down and do that on their own without a teacher no matter how much input I give to them before the Good Behaviour Game... if it’s an independent task and it’s writing or a difficult numeracy task then I’d say they do that better outside of The Good Behaviour Game.” (school 6, teacher D)*

Of particular interest given our findings in relation to H2 (the prediction that the GBG would produce amplified effects for boys at-risk for developing conduct problems), another teacher commented that they had an issue with two boys with behaviour plans, as *“trying to engage those two boys in particular, was really tricky”* (school 1, teacher H), but that as time went on, they noticed, *“he’s really turned it round he’s been a little superstar recently... he quickly proved himself and he’s already I’ve put him as a team leader now because he’s responsible...so it’s really having an effect on him”* (school 1, teacher H). Another teacher agreed, finding *“it works really well especially with my boys who can be often quite destructive, and hard work”* (school 2, teacher M).

However, other teachers felt that the GBG benefited all pupils *“equally”* (school 1, teacher E), and *“when I think about the whole class in general, the boys have stuck to it and then even in terms of sort of attainment with highers, middles, lowers, they all have responded in a similar way”* (school 1, teacher H). This was particularly evident towards the end of the trial, as teachers who had initially noted differential gains felt that while there were *“certain pupils that respond better than others...I don’t think I can say if it’s a certain group of children really, no”* (school 4, j).

## Sustainability

At the end of the trial, some teachers expressed a preference to continue implementing the GBG, especially if they were going to be teaching pupils who were already familiar with it: *“if I’m still in Year Four I think it’s a possibility that I would continue because I do know that the Year Threes...have played the game on a regular basis”*<sup>17</sup> (school 6, teacher K). However, they did acknowledge that the significant effort and time to introduce the GBG to a new class of children: *“you need to go through building up playing the game, but it again it boils down to time and constraints of the timetable and fitting it in”* (school 1, teacher E).

Pupils’ familiarity with the GBG seemed to be a key motivation for SLT members to widen the participation in the intervention across the school, particularly in the lower years, where the emphasis on independence during gameplay was seen to be beneficial:

---

<sup>17</sup> Some Year 3 teachers continued to implement the GBG in the second year of the trial with pupils in the year below the target cohort.

*"We will probably take it down to Year Two I would think if we if we could sort of try it because I just think the earlier children realise that they're more able to do on their own the better they will be... the sooner you realise that actually you can learn yourself you don't need an adult to be propping you up all the while...progress and interest in learning takes off for children 'cos they do realise they can do it themselves." (school 4, Head Teacher)*

SLT endorsement of a wider roll out appeared to be driven by teacher feedback, *"if we're seeing the benefits of it, and the classroom teachers are telling me we are... I'd like it to move across the school so we're all buying into it as it's part of our culture that that's what we do."* (school 2, GBG co-ordinator), alongside a strategic preference for the continuity and consistency offered by whole school approaches: *"I would love to see it as a whole school approach and I think even if it followed children up to high school as well then it's definitely going to make such an impact"* (school 6, teacher D). Some SLT members were particularly keen on introducing GBG to other year groups in order to target classes that were known to be particularly challenging: *"the frustration from my point is I can see that it's working and I'd like to trial it...in other classes in other year groups to see the impact but obviously that's only going to filter through when every member of staff [has] had the training"* (school 6, Head Teacher).

Maintaining procedural fidelity was noted as an important aspect of sustained implementation, with one Head Teacher stating that, *"we'd probably keep [GBG] more or less as it is because we think that if you're going to use it then you must use it in the way it is formatted or else you're not really comparing like with like are you?"* (school 4, Head Teacher). However, while teachers expressed similar sentiments with regard to certain aspects, such as the four GBG rules, *"I think I'd certainly like to stick with the four rules 'cos I think that really simplifies the rules you're asking the children to follow"* (school 3, teacher B), they also noted that continued delivery of the intervention would require adaptations to meet needs of the pupils in their class:

*"I'd stick with playing a game and using those scoreboards when I feel necessary...but I think that we would possibly tweak some of the actual rules of the game play in terms of how we can interact...if you need an adult to support a group with their learning, sort of an SEN table, that adult has to become a player in the game so they can't actually behave like an adult in the class and that can sometimes limit what that adult can do so I think in those areas we'd tweak it to sort of match how we need it to be."* (school 3, teacher B)

## Conclusion

### Key conclusions

1. We found no evidence that the GBG improves pupils' reading. This result has a high security rating.
2. We found no evidence that the GBG improves pupils' behaviour (specifically, concentration problems, disruptive behaviour, and pro-social behaviour).
3. Implementation was variable and in particular, the frequency and duration with which the GBG was played did not reach the levels expected by the developer. One-quarter of schools in the intervention arm ceased implementation before the end of the trial.
4. Higher levels of pupil engagement with the game were associated with improved reading, concentration, and disruptive behaviour scores at follow-up. There was no clear evidence that other aspects of implementation (for example, how well or how frequently the game was played) were related to whether pupil outcomes improved. These results were sensitive to changes in how we analysed the data, and so should be interpreted with caution.
5. There was tentative evidence that boys identified as at-risk of developing conduct problems at the beginning of the project benefitted from the GBG. For these children, small reductions in concentration problems and disruptive behaviour were observed.

### Interpretation

The current study is, to the best of our knowledge, the largest RCT of the GBG conducted worldwide to date. It is the first trial of the intervention in the UK, and among the first internationally to examine academic outcomes (in this case, reading scores) using a design in which effects are tested in isolation (e.g. not in combination with an academically-focused intervention, as in Bradshaw et al., 2009). Taken together, our findings and those of Kellam et al. (2008) and Dion et al. (2011) are consistent in identifying no impact on attainment at the ITT level. Of course, this does not rule out longer-term, so-called 'sleeping' effects, or indeed subgroup effects (points which we address below), but in the short-term the preponderance of evidence suggests that the GBG does not improve pupils' academic performance. However, we also found no ITT effect on their disruptive behaviour, concentration problems, or pro-social behaviour. At first, this seems incongruent with the weight of the evidence base outlined in the Introduction chapter, until one begins to interrogate its nature. So, for example, while Dolan et al. (1993) reported small, proximal effects on aggressive and shy behaviours at the ITT level in the original GBG RCT in Baltimore, Maryland, the study design was very different to the current trial (e.g. less than 400 pupils across 20 classrooms in the GBG versus external control comparisons; sample predominantly black and ethnic minority students). Perhaps more importantly, teachers in the Baltimore trial received significantly more training (40 hours) and implemented the GBG over a much shorter period of time (less than one school year; outcomes assessed in the autumn and spring terms).

In another example, Leflot, van Lier, Onghena, and Colpin's (2010) RCT in Belgium found small to moderate proximal effects of the GBG in their ITT analysis with regard to pupils' on-task, talking out, and oppositional behaviours. Again though, study design differences abound (e.g. within-school randomisation; less than 600 children across 20 classrooms), and though they provide no data, the authors reported that all but one of the teachers in the intervention group reached the recommended levels of implementation in terms of frequency and duration of gameplay by the end of the trial period. Such differences may plausibly account for the divergence of findings observed. More broadly, the issue of cultural transferability (or lack thereof) should not be ignored (Weare & Nind, 2011). It may simply be that certain principles and practice inherent in the GBG are incompatible with the English cultural

context in education, and in particular the preferred practices of teachers. Indeed, our qualitative IPE data supports this assertion; recall that many teachers struggled with certain mandated intervention procedures, most notably not being able to directly interact or intervene with pupils during gameplay. However, we also note Greenberg's (2010) salient point about the 'prevention paradox': "In universal interventions, it is usually the case that a large percentage of the population begins without symptoms and thus it is unlikely (at least in the short term) that much of this population will change" (p.34). This observation is supported by our baseline outcome data, which demonstrated high levels of pro-social behaviour, low levels of disruptive behaviour, and only moderate levels of concentration problems in the trial sample, leaving minimal room for change at the ITT level.

Our findings in relation to H2, although tentative given the marginal nature of the effects identified (and sensitivity to changes in modelling parameters), are consistent with the proposition that the GBG is particularly effective for boys at-risk of developing conduct problems. This subgroup effect is in keeping with theory (e.g. gendered socialisation of competitiveness, Gneezy, Leonard, & List, 2009; responses to social task demands in the classroom, Kellam et al., 1994), consistent with the earlier findings (e.g. Kellam, Rebok, Jolongo, & Mayer, 1994), and aligns with the views of some teachers ascertained in the qualitative case study strand of the IPE. It can be argued that the intervention ESs observed in relation to disruptive behaviour ( $g=-0.30$ ) and concentration problems ( $g=-0.29$ ) are practically significant and meaningful when one considers that they were achieved through the provision of a relatively low-intensity, universal intervention, and the fact that even modest decreases in behavioural problems among at-risk children can have significant consequences for the broader school environment (Deighton et al., 2015). Furthermore, there is arguably much value in an intervention that can potentially moderate maladaptive developmental trajectories that yield such huge personal and societal costs later in life. For example, childhood conduct problems among boys are associated with a two- to threefold increase in costs by early adulthood, driven primarily by criminal justice contacts (D'Amico et al., 2014). Clearly, the findings of this trial are promising rather than definitive as the effects observed are only in the short-term (e.g. immediately post-intervention). Longer-term follow-up is warranted in order to determine whether they are sustained or attenuate over time. Such longer-term follow-up would also allow the identification of sleeper effects on attainment in this subgroup, which would be in line with our recent findings in relation to the erosive effects of early conduct problems on later attainment among boys (Panayiotou & Humphrey, 2017); the GBG may be an effective means through which to disrupt these negative developmental cascade effects. The evaluation team are in the process of exploring this, courtesy of a follow-up grant (see 'Future research and publications' below).

In contrast to the above, our analyses pertaining to H3 yielded no evidence of differential effects of the GBG among children eligible for FSM, for any pupil-level outcome. However, this was very much an exploratory subgroup analysis, based on the premise that school-based interventions may compensate for some of the factors that constrain the achievement of pupils from socio-economically deprived backgrounds (Dietrichson et al., 2017), alongside provisional evidence that other universal, preventive approaches can produce differential gains of this nature (Holsen et al., 2009). Whilst several earlier GBG trials have assessed SES in some way, none have yet examined whether intervention effects vary as a function of poverty at the individual-level. For example, Weis et al. (2015) found amplified effects of the GBG among students attending low and moderate (compared to high) SES *schools*, but their analysis did not extend to variability in SES among *pupils* attending said schools. Thus, unlike the above hypotheses, there is no immediate frame of reference for our findings in relation to the FSM subgroup. Nevertheless, based on the comprehensive and consistent pattern of null findings, including in our sensitivity analyses, we feel it safe to conclude that the GBG is equally ineffective in improving attainment for pupils from socio-economically deprived backgrounds and their more affluent peers. Furthermore, while it is impossible to rule out the emergence of sleeper effects among the FSM subgroup, this seems unlikely. Compared to the at-risk boys subgroup, in which short-term effects on concentration and behaviour were identified which could feasibly trigger longer-term benefits in academic attainment, there was no such pattern here. This is compounded by the fact that, despite

theorising *potential* benefits for children from low SES backgrounds (e.g. “*Our school has a certain level of.... deprivation... some of those factors... can bring themselves into the school so we thought it would be a good idea to give them some structure to help them modify their own behaviours and that [the GBG] seemed to fit that sort of ethos*”; school 2, GBG co-ordinator) and subsequently signposting perceived differential gains among a number of other groups (e.g. pupils with SEN, varying levels of attainment, and/or boys with behavioural problems), teachers in case study schools provided no indication that they felt that those eligible for FSM experienced any distinct benefit from participation in the GBG.

With regard to H4, our IPE data provided very useful insights that help to contextualise and explain our impact findings. Descriptive implementation data gathered from the online scoreboard revealed that the GBG was played less frequently and for shorter periods of time than recommended by the developer. However, among the very few GBG studies in which dosage has been reported, lower than expected levels appear to be the norm (e.g. lower than expected frequency reported by Hagermoser-Sanetti & Fallon, 2011; lower than expected duration in Domitrovich et al., 2015). We also note that our implementation-outcomes analyses pertaining to dosage were mixed at best; thus, *more* of the GBG did not consistently equate to *improved* outcomes. When the game was played, our observational data suggested that teachers followed most of the prescribed procedures associated with the game and did so in an enthusiastic, engaging manner. Almost all children in a given class were present when the GBG was played, and they responded favourably, albeit with a drop in the second year of the trial. While no other studies have focused on reach and responsiveness to date, several have examined fidelity, and here our findings are well within the range identified, albeit towards the lower end (e.g. much higher fidelity than in Ialongo et al., 2001; but lower than in Domitrovich et al., 2015; Dion et al., 2011; Leflot et al., 2013).

As above though, our analyses did not support the assertion that higher fidelity was associated with improved outcomes; quite the opposite, in fact. Indeed, the only implementation dimension where higher levels were consistently associated with improved outcomes was participant responsiveness. This pattern of findings suggests that the behaviour of the recipients of the GBG is as important (or even more important) than that of the implementer in terms of later outcomes. Specifically, pupil-level reading, concentration problems and disruptive behaviour scores at follow-up were significantly improved in classes where pupils were more responsive (for example, being attentive to their teacher’s instructions, more frequently correcting their behaviour following an infraction, and showing greater interest in rewards). This raises the possibility that the null impact findings outlined above may be underpinned by a failure in the overall trial sample to internalise and subsequently generalise learned behaviours to contexts beyond the game. Scrutiny of the online scoreboard data in Figure 3 aligns with this proposition. Teams routinely won the game, even from the outset, but there was no clear reduction in infractions over time observed in probe sessions (in stark contrast to the Oxfordshire pilot, Chan et al, 2012). Pennington and McComas’ (2017) recent small-scale GBG study provides further support, noting a failure to generalise learned behaviour across different classroom contexts.

Our usual provision analyses are also particularly important for interpreting our impact findings, as they suggest that programme differentiation may have been too limited for the GBG to produce meaningful changes in our primary and secondary outcomes. Recall that teachers in both trial arms reported frequent use of several core aspects of the GBG at baseline as part of their existing approaches to behaviour management (e.g. classroom rules, observation and monitoring of pupil behaviour, provision of rewards, use of groups). In relation to the intervention arm, the GBG may therefore have been insufficiently distinct from what teachers were already doing. This in turn may explain the lower than expected dosage findings, if teachers felt that their usual practices were sufficient (or sufficiently similar). With regard to the usual provision arm, our data suggests that the trial counterfactual was remarkably similar – in broad terms, at least – to the intervention being tested. Under such conditions, the meagre impact findings are arguably unsurprising. Finally, our usual provision findings may also link

to the theme of cultural transferability noted above. The perception of need for effective classroom management is probably no different across the various countries and cultures in which the GBG has been studied, but the processes through which it is usually achieved may vary considerably. Perhaps greater impact has been seen in these other countries because their usual provision differs from that seen here.

In light of the various findings discussed above, it is unsurprising that our analyses pertaining to H5 failed to identify any impact on teacher-level outcomes. Drawing on the pro-social classroom model (Jennings & Greenberg, 2009), we proposed that the reciprocal inter-relations between teacher wellbeing, classroom-level processes, and pupil outcomes, could lead to improved teacher outcomes as a result of participation in the GBG. However, the predicted effects of reduced stress, and improved self-efficacy in classroom management and retention, were all essentially theorised as *secondary* to expected changes in pupil-level outcomes. For example, we predicted that improvements in teachers' self-efficacy in classroom management could eventuate *if* the GBG was implemented well and produced observable changes in pupil behaviour. Given that neither of these conditions were fully met in the trial, the likelihood of changes in teacher-level outcomes was very low. A further possibility is that the additional demands of a new intervention (and participation in a research study) meant that there was no reduction in stress seen at this stage. It is possible that the GBG has to be implemented for longer and become more embedded for reduction in teacher stress levels to be observed.

### Strengths and limitations

This study has numerous strengths, increasing our confidence that the principal (impact) findings are secure. We utilised a cluster-randomised design with appropriate analysis that took account of the hierarchical and clustered nature of the dataset. The trial was large and well powered, with an MDES of 0.13 at randomisation. Attrition was 0% at the school-level and 19% at the pupil-level. There was no evidence of differential attrition between trial arms in our analysis of missing data. Balance on observables in the analysis sample was good, with negligible differences between pupil-level outcomes at baseline. In terms of threats to internal validity, the intervention being trialled has been outlined in detail using the recommended TIDieR framework. The use of cluster-randomisation minimised the possibility of diffusion/contamination, as did the supply control of intervention materials, training and support by the delivery team (Mentor UK). There was no evidence of compensation rivalry or resentful demoralisation in the usual provision group. The outcome measures used were evaluated by the EEF and considered to be reliable, externally valid and not intervention specific (e.g. not 'inherent to treatment'). Randomisation was conducted independently of the evaluation team by a statistician in the MAHSC-CTU. All analysis and reporting is based on CONSORT standards and EEF guidelines, and a range of pre-specified sensitivity analyses (e.g. MI of missing data, inclusion of minimisation variables at the school level) were undertaken. Our principal findings relating to the impact of the GBG on pupil-level outcomes at the ITT level (H1) were not sensitive to any changes in our modeling parameters. Our findings in relation to planned subgroup analyses (H2, H3) were also largely insensitive to such changes. In terms of generalizability, the trial was conducted in 23 Local Authorities across the three regions from which we recruited schools. Trial school composition mirrored that of primary schools in England in respect of size and the proportion of students speaking EAL, but contained significantly larger proportions of children with SEND and eligible for FSM, in addition to lower rates of absence and attainment. However, this is arguably typical for trials of this nature, wherein recruitment naturally skews towards schools with such profiles (Humphrey et al., 2015).

Nevertheless, a number of limitations and complicating factors also need to be considered. Firstly, as noted in the Method chapter, approximately one-quarter (n=9, 24%) of schools in the intervention arm formally discontinued their implementation of the GBG during the trial. It is possible that the lack of impact identified in the current study is underpinned by this implementation failure. However, it is our contention that the pattern observed here is likely a 'best case' scenario for real world practice, given

the resource made available through the project funding to optimise implementation (e.g. subsidised intervention cost for schools, additional provision for GBG data monitoring made available by the evaluation team, developer support for the delivery team).

Furthermore, and perhaps most tellingly, reanalysis of the main trial data (not reported here, but available on request from the authors) in which we excluded the nine schools that discontinued implementation from the intervention arm revealed no substantive difference in our main findings. Secondly, in respect of our secondary outcome measures, we note that the use of direct observational measures of pupil behaviour are generally considered to be more robust than informant-report surveys such as the TOCA-C (Merrell, 2008). However, this was not feasible given the constraints of the current trial (e.g. financial and human resource; data burden), and in any event would likely have produced further complications that would undermine the relative advantages of the method (e.g. as noted earlier, blinding of researchers would have been impossible given the number of physical artifacts associated with the GBG). Thirdly, in observing each teacher implementing the GBG once, we were only able to provide a snapshot (albeit a very useful one) of implementation. While researcher/observer effects are not a major concern (after all, teachers' practice is routinely observed, and indeed they were observed delivering the GBG on multiple occasions by their coaches), multiple independent observations over time would have allowed temporal patterns in implementation to be identified and taken into account in analyses (Humphrey et al., 2016). However, as a counterpoint, several studies have demonstrated remarkable stability in implementation, suggesting minimal added value in multiple observations (Humphrey, Barlow & Lendrum, 2017). Finally, in taking a 'variable-focused' approach to our analysis for H4, we neglected to consider how different dimensions of implementation may interact to moderate outcomes. Alternative methodologies involving the use of 'person-focused' modeling (e.g. through identification of latent implementation profiles among teachers, as in Low, Smolkowski, & Cook, 2016) may have produced different findings.

### **Future research and publications**

Shortly after being commissioned, this EEF evaluation grant was supplemented by funding from the National Institute for Health Research (NIHR). The NIHR extension grant is being used to determine the impact of the GBG on health-related outcomes for children (e.g. mental health, health-related quality of life), the sustainability (or emergence) of effects on said outcomes (in addition to those captured in the EEF project) at one- and two-year follow-up, and the extent to which the intervention can be regarded as cost-effective. This enables us to pursue several key avenues of inquiry triggered by the findings of the current trial (e.g. is the subgroup effect noted for boys at-risk of developing conduct problems maintained? Are there sleeper effects on academic attainment?). These questions are particularly pertinent in light of the GBG evidence base. For example, Kellam, Rebok, Jaffee, and Mayer's (1994) study revealed that the main proximal effects of the GBG on behavioural outcomes had attenuated at long-term follow-up, but subgroup effects identified for aggressive boys persisted. In terms of academic outcomes, Hemelt, Roth, and Eaton's (2013) long-term follow-up of Dolan et al.'s (1993) trial sample revealed a similar picture in terms of null ITT findings, but also found that subgroup effects emerged in relation to gender (though intriguingly, and in contradiction to what might be theorised, it appeared that females benefited more than males). Clearly there is much scope for further inquiry on longer-term outcomes, especially given the findings of a recent systematic review of a range of interventions designed to prevent behavioural problems (including the GBG), which found limited evidence for effects beyond 6 months post-intervention (Smedler, Hjern, Wiklund, Anttila, & Pettersson, 2015).

Further research is also underway that makes use of the data collected in this trial. Emma Ashworth is in the process of completing a doctoral thesis drawing on cumulative risk theory to examine differential gains associated with the GBG among pupils at varying levels of risk exposure, and to explore the extent to which any differential gains are moderated by implementation variability (e.g. do pupils



exposed to high levels of cumulative risk benefit more from higher GBG dosage than those exposed to lower levels of risk?). Kirsty Frearson is in the process of completing a doctoral thesis that uses social disorganisation theory to examine whether differential gains among pupils eligible for free-school meals vary as a function of school-level poverty, the extent to which school-level poverty predicts teachers' implementation behaviour, and whether said behaviour can be classified using person-focused approaches in order to predict pupil-level outcomes.

## References

- Bandura, A. (1986). *Social foundations of thought and action: A social cognitive 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. *Journal of Applied Behavior Analysis*, 2, 119–24.
- Berkel, C., Mauricio, A. M., Schoenfelder, E., & Sandler, I. N. (2011). Putting the pieces together: an integrated model of program implementation. *Prevention Science*, 12, 23–33. <http://doi.org/10.1007/s11121-010-0186-1>
- Borg, M. G., Riding, R. J. & Falzon, J. M. (1991). Stress in teaching: A study of occupational stress and its determinants, job satisfaction and career commitment among primary schoolteachers. *Educational Psychology*, 11, 59-75.
- Boyle, G. J., Borg, M. G., Falzon, J. M., & Baglioni, A. J. (1995). A structural model of the dimensions of teacher stress. *The British Journal of Educational Psychology*, 65, 49–67.
- Bradshaw, C. P., Zmuda, J. H., Kellam, S. G., & Ialongo, N. S. (2009). Longitudinal impact of two universal preventive interventions in first grade on educational outcomes in High School. *Journal of Educational Psychology*, 101, 926–937. <http://doi.org/10.1037/a0016586>
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3, 77–101. <http://doi.org/10.1191/1478088706qp063oa>
- Breeman, L. D., van Lier, P. A. C., Wubbels, T., Verhulst, F. C., van der Ende, J., Maras, A., ... 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, 156–167. <http://doi.org/10.1177/1098300715593466>
- Carpenter, J. J. R., Goldstein, H., & Kenward, M. G. M. (2011). REALCOM-IMPUTE software for multilevel multiple imputation with mixed response types. *Journal of Statistical Software*, 45, 1–12. <http://doi.org/http://dx.doi.org/10.18637/jss.v045.i05>
- Chan, G., Foxcroft, D., Smurthwaite, B., Coombes, L., & Allen, D. (2012). *Improving child behaviour management: An evaluation of the Good Behaviour Game in UK primary schools*. Oxford: Oxford Brookes University.
- 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, 369–387. <http://doi.org/10.1002/casp.2268>
- D’Amico, F., Knapp, M., Beecham, J., Sandberg, S., Taylor, E., & Sayal, K. (2014). Use of services and associated costs for young adults with childhood hyperactivity/conduct problems: 20-year follow-up. *British Journal of Psychiatry*, 204, 441–7. <http://doi.org/10.1192/bjp.bp.113.131367>
- Deighton, J., Humphrey, N., Wolpert, M., Patalay, P., Belsky, J., & Vostanis, P. (2015). An evaluation of the implementation and impact of England’s mandated school-based mental health initiative in elementary schools. *School Psychology Review*, 44, 117-138. <https://doi.org/10.17105/SPR44-1.117-138>
- Department for Education. (2012a). *Ensuring good behaviour in schools*. London: DfE.
- Department for Education. (2012b). *Pupil behaviour in schools in England*. London:DfE.
- Department for Education. (2014a). *Behaviour and discipline in schools*. London:DfE.
- Department for Education. (2014b). *School behaviour and attendance: research priorities and*

- questions*. London: DfE.
- Department for Education. (2015). *Schools, pupils and their characteristics: January 2015*. London: DfE.
- Department for Education. (2015) *Special educational needs in England: January 2015*. London: DfE.
- Department for Education. (2015). *National curriculum assessments at key stage 2 in England, 2015 (revised)*. London: DfE.
- Department for Education. (2016). *Mental health and behaviour in schools*. London: DfE.
- Department for Education. (2016). *Pupil absence in schools in England: 2014 to 2015*. London: DfE.
- Department for Education. (2017). *Analysis of school and teacher level factors relating to teacher supply*. London: DfE.
- Department for Education. (2017). *Schools, pupils and their characteristics: January 2017*. London: DfE.
- Dietrichson, J., Bøg, M., Filges, T., & Klint Jørgensen, A.-M. (2017). Academic interventions for elementary and middle school students with low socioeconomic status: A systematic review and meta-analysis. *Review of Educational Research, 87*, 243–282. <http://doi.org/10.3102/0034654316687036>
- Dijkman, M. A. M., Harting, J., & van der Wal, M. F. (2015). Adoption of the Good Behaviour Game: An evidence-based intervention for the prevention of behaviour problems. *Health Education Journal, 74*, 168–182. <http://doi.org/10.1177/0017896914522234>
- Dion, E., Roux, C., Landry, D., Fuchs, D., Wehby, J., & Dupéré, V. (2011). Improving attention and preventing reading difficulties among low-income first-graders: A randomized study. *Prevention Science, 12*, 70–79. <http://doi.org/10.1007/s11121-010-0182-5>
- Dolan, L. J., Kellam, S. G., Brown, C. H., Werthamer-Larsson, L., Rebok, G. W., Mayer, L. S., ... Wheeler, L. (1993). The short-term impact of two classroom-based preventive interventions on aggressive and shy behaviors and poor achievement. *Journal of Applied Developmental Psychology, 14*, 317–345. [http://doi.org/10.1016/0193-3973\(93\)90013-L](http://doi.org/10.1016/0193-3973(93)90013-L)
- Domitrovich, C. E., Bradshaw, C. P., Berg, J. K., Pas, E. T., Becker, K. D., Musci, R., ... Jalongo, N. (2016). How so school-based prevention programs impact teachers? Findings from a randomized trial of an integrated classroom management and social-emotional program. *Prevention Science, 17*, 325–337. <http://doi.org/10.1007/s11121-015-0618-z>
- Domitrovich, C. E., Bradshaw, C. P., Greenberg, M. T., Embry, D., Poduska, J. M., & Jalongo, N. S. (2010). Integrated models of school-based prevention: Logic and theory. *Psychology in the Schools, 47*, 71–88. Retrieved from <http://doi.wiley.com/10.1002/pits.20452>
- Domitrovich, C. E., Pas, E. T., Bradshaw, C. P., Becker, K. D., Keperling, J. P., Embry, D. D., & Jalongo, N. (2015). Individual and school organizational factors that influence implementation of the PAX Good Behavior Game intervention. *Prevention Science, 16*, 1064–1074. <http://doi.org/10.1007/s11121-015-0557-8>
- Donaldson, J. M., & Wiskow, K. M. (2017). The Good Behaviour Game. In B. Teasdale & M. S. Bradley (Eds.), *Preventing Crime and Violence* (pp. 229–241). Springer: Switzerland. <http://doi.org/10.1007/978-3-319-44124-5>
- Durlak, J. A. (2016). Programme implementation in social and emotional learning: basic issues and research findings. *Cambridge Journal of Education, 46*, 333-345. <http://doi.org/10.1080/0305764X.2016.1142504>
- Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: a review of research on the influence of

- implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology*, 41, 327–50. <http://doi.org/10.1007/s10464-008-9165-0>
- Elswick, S., & Casey, L. (2011). The good behavior game is no longer just an effective intervention for students: An examination of the reciprocal effects on teacher behaviors. *Beyond Behavior*, 21, 36–46.
- Embry, D. D. (2002). The Good Behavior Game: a best practice candidate as a universal behavioral vaccine. *Clinical Child and Family Psychology Review*, 5, 273–97.
- Farrell, A. D., Henry, D. B., & Bettencourt, A. (2013). Methodological challenges examining subgroup differences: Examples from universal school-based youth violence prevention trials. *Prevention Science*, 14, 121–133. <http://doi.org/10.1007/s11121-011-0200-2>
- Fishbein, J. E., & Wasik, B. H. (1981). Effect of the Good Behavior Game on disruptive library behavior. *Journal of Applied Behavior Analysis*, 14, 89–93. <http://doi.org/10.1901/jaba.1981.14-89>
- Fixsen, D., Naoom, S. F., Blase, K. A., Friedman, R. M., & Wallace, F. (2005). *Implementation research: a synthesis of the literature*. Tampa: University of South Florida.
- Flower, A., McKenna, J. W., Bunuan, R. L., Muething, C. S., & Vega, R. (2014). Effects of the Good Behavior Game on challenging behaviors in school settings. *Review of Educational Research*, 84, 546–571. <http://doi.org/10.3102/0034654314536781>
- Ford, C., Keegan, N., Poduska, J., Kellam, S., & Littman, J. (2014). *Good Behaviour Game implementation manual*. Washington, DC: American Institutes for Research.
- Forman, S., Olin, S., Hoagwood, K., & Crowe, M. (2009). Evidence-based interventions in schools: developers' views of implementation barriers and facilitators. *School Mental Health*, 1, 26–36. <http://doi.org/10.1007/s12310-008-9002-5>
- Gneezy, U., Leonard, K. L., & List, J. A. (2009). Gender differences in competition: Evidence from a matrilineal and a patriarchal society. *Econometrica*, 77, 1637–1664.
- Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. *Journal of Child Psychology and Psychiatry*, 38, 581–586. <http://doi.org/10.1111/j.1469-7610.1997.tb01545.x>
- Greenberg, M., Domitrovich, C., Graczyk, P., Zins, J., & Services, C. for M. H. (2005). *The study of implementation in school-based preventive interventions: Theory, research, and practise*. Rockville: CMHS.
- Greenberg, M. T. (2010). School-based prevention: current status and future challenges. *Effective Education*, 2, 27–52.
- Greenberg, M. T., & Abenavoli, R. (2017). Universal interventions: Fully exploring their impacts and potential to produce population-level impacts. *Journal of Research on Educational Effectiveness*, 10, 40–67. <http://doi.org/10.1080/19345747.2016.1246632>
- Grierson, J. (2017). Behaviour is a national problem in schools in England, review finds. *The Guardian*, March 24.
- Gupta, S. K. (2011). Intention-to-treat concept: A review. *Perspectives in Clinical Research*, 2, 109–12. <http://doi.org/10.4103/2229-3485.83221>
- Gutman, L. M., & Vorhaus, J. (2012). *The impact of pupil behaviour and wellbeing on educational outcomes*. London: DFE.
- Hagermoser Sanetti, L. M., & Fallon, L. M. (2011). Treatment integrity assessment: How estimates of adherence, quality, and exposure influence interpretation of implementation. *Journal of Educational and Psychological Consultation*, 21, 209–232.

<http://doi.org/10.1080/10474412.2011.595163>

- Hansen, W. (2014). Measuring fidelity. In Z. Sloboda & H. Petras (Eds.), *Defining prevention science* (pp. 335–359). New York, NY: Springer.
- Haydn, T. (2014). To what extent is behaviour a problem in English schools? Exploring the scale and prevalence of deficits in classroom climate. *Review of Education, 2*, 31–64. <http://doi.org/10.1002/rev3.3025>
- Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics, 32*, 341–370. <http://doi.org/10.3102/1076998606298043>
- Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis, 29*, 60–87. <http://doi.org/10.3102/0162373707299706>
- Hemelt, S. W., Roth, K. B., & Eaton, W. W. (2013). Elementary school interventions: Experimental evidence on postsecondary Outcomes. *Educational Evaluation and Policy Analysis, 35*, 413–436. <http://doi.org/10.3102/0162373713493131>
- Hoffmann, T. C., Glasziou, P. P., Boutron, I., Milne, R., Perera, R., Moher, D., ... Michie, S. (2014). Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. *BMJ, 348*, g1687. <http://doi.org/10.1136/bmj.g1687>
- Holsen, I., Iversen, A. C., & Smith, B. H. (2009). Universal social competence promotion programme in school: Does it work for children with low socio-economic background? *Advances in School Mental Health Promotion, 2*, 51–60. <http://doi.org/10.1080/1754730X.2009.9715704>
- Humphrey, N., Barlow, A., & Lendrum, A. (2017). Quality matters: Implementation moderates student outcomes in the PATHS Curriculum. *Prevention Science, Early View*, 1–12. <http://doi.org/10.1007/s11121-017-0802-4>
- Humphrey, N., Barlow, A., Wigelsworth, M., Lendrum, A., Pert, K., Joyce, C., ... Turner, A. (2015). *Promoting Alternative Thinking Strategies (PATHS): Evaluation report*. London: Education Endowment Foundation.
- Humphrey, N., Lendrum, A., Ashworth, E., Frearson, K., Buck, R., & Kerr, K. (2016). *Implementation and process evaluation (IPE) for interventions in educational settings: an introductory handbook*. London: Education Endowment Foundation.
- Ialongo, N., Poduska, J., Werthamer, L., & Kellam, S. (2001). The distal impact of two first-grade preventive interventions on conduct problems and disorder in early adolescence. *Journal of Emotional and Behavioral Disorders, 9*, 146–160. <http://doi.org/10.1177/106342660100900301>
- Ialongo, N. S., Werthamer, L., Kellam, S. G., Brown, C. H., Wang, S., & Lin, Y. (1999). Proximal impact of two first-grade preventive interventions on the early risk behaviors for later substance abuse, depression, and antisocial behavior. *American Journal of Community Psychology, 27*, 599–641. <http://doi.org/10.1023/A:1022137920532>
- Jennings, P. A., & Greenberg, M. T. (2009). The prosocial classroom: Teacher social and emotional competence in relation to student and classroom outcomes. *Review of Educational Research, 79*, 491–525. <http://doi.org/10.3102/0034654308325693>
- 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., Ling, X., Merisca, R., Brown, C. H., & Ialongo, N. S. (1998). The effect of the level of aggression in the first grade classroom on the course and malleability of aggressive behavior into middle school. *Development and Psychopathology, 10*, 165–185.

<http://doi.org/10.1017/S0954579498001564>

- 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, 73–84.
- Kellam, S. G., Rebok, G. W., Ialongo, N., & Mayer, L. S. (1994). The course and malleability of aggressive behavior from early first grade into middle school: results of a developmental epidemiologically-based preventive trial. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, 35, 259–81.
- Kelm, J. L., & McIntosh, K. (2012). Effects of school-wide positive behavior support on teacher self-efficacy. *Psychology in the Schools*, 49, 137–147. <http://doi.org/10.1002/pits.20624>
- Kleinman, K. E., & Saigh, P. A. (2011). The effects of the good behavior game on the conduct of regular education new york city high school students. *Behavior Modification*, 35, 95–105. <http://doi.org/10.1177/0145445510392213>
- Korpershoek, H., Harms, T., de Boer, H., van Kuijk, M., & Doolaard, S. (2016). A meta-analysis of the effects of classroom management strategies and classroom management programs on students' academic, behavioral, emotional, and motivational outcomes. *Review of Educational Research*, 86, 1–38. <http://doi.org/10.3102/0034654315626799>
- Koth, C. W., Bradshaw, C. P., & Leaf, P. J. (2009). Teacher Observation of Classroom Adaptation Checklist: Development and factor structure. *Measurement and Evaluation in Counseling and Development*, 42, 15–30. <http://doi.org/10.1177/0748175609333560>
- Lannie, A. L., & McCurdy, B. L. (2007). Preventing disruptive behavior in the urban classroom: Effects of the Good Behavior Game on student and teacher Behavior. *Education and Treatment of Children*, 30, 85–98. <http://doi.org/10.1353/etc.2007.0002>
- Leflot, G., van Lier, P. a C., Onghena, P., & Colpin, H. (2013). The role of children's on-task behavior in the prevention of aggressive behavior development and peer rejection: A randomized controlled study of the Good Behavior Game in Belgian elementary classrooms. *Journal of School Psychology*, 51, 187–199. <http://doi.org/10.1016/j.jsp.2012.12.006>
- Leflot, G., van Lier, P., 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, 869–82. <http://doi.org/10.1007/s10802-010-9411-4>
- Lendrum, A., & Humphrey, N. (2012). The importance of studying the implementation of school-based interventions. *Oxford Review of Education*, 38, 635–652.
- Lightfoot, L. (2016). Nearly half of England's teachers plan to leave in next five years. *The Guardian*, March 22.
- Low, S., Smolkowski, K., & Cook, C. (2016). What constitutes high-quality implementation of SEL programs? A latent class analysis of Second Step® implementation. *Prevention Science*, 17, 981–991. <http://doi.org/10.1007/s11121-016-0670-3>
- Lynch, D., & Keenan, M. (2016). The good behaviour game: Maintenance effects. *International Journal of Educational Research*, (2015), *Early View*, 1–9. <http://doi.org/10.1016/j.ijer.2016.05.005>
- Lynne, S. (2015). *Investigating the use of a positive variation of the Good Behavior Game in a high school setting*. Mississippi: University of Southern Mississippi.
- McCormick, J., & Barnett, K. (2011). Teachers' attributions for stress and their relationships with burnout. *International Journal of Educational Management*, 25, 278–293. <http://doi.org/10.1108/09513541111120114>

- McCurdy, B. L., Lannie, A. L., & Barnabas, E. (2009). Reducing disruptive behavior in an urban school cafeteria: An extension of the Good Behavior Game. *Journal of School Psychology, 47*, 39–54. <http://doi.org/10.1016/j.jsp.2008.09.003>
- McGoey, K. E., Schneider, D. L., Rezzetano, K. M., Prodan, T., & Tankersley, M. (2010). Classwide intervention to manage disruptive behavior in the kindergarten classroom. *Journal of Applied School Psychology, 26*, 247–261. <http://doi.org/10.1080/15377903.2010.495916>
- Merrell, K. W. (2008). *Behavioural, social, and emotional assessment of children and adolescents (3rd Edition)*. Oxon: Routledge.
- Moore, J. E., Bumbarger, B. K., & Cooper, B. R. (2013). Examining adaptations of evidence-based programs in natural contexts. *The Journal of Primary Prevention, 34*, 147–61. <http://doi.org/10.1007/s10935-013-0303-6>
- National Institute for Health and Care Excellence. (2013). *Antisocial behaviour and conduct disorders in children and young people: recognition, intervention and management*. London: NICE.
- Nolan, J. D., Filter, K. J., & Houlihan, D. (2014). Preliminary report: An application of the good behavior game in the developing nation of Belize. *School Psychology International, 35*, 421–428. <http://doi.org/10.1177/0143034313498958>
- Nolan, J. D., Houlihan, D., Wanzek, M., & Jenson, W. R. (2014). The Good Behavior Game: A classroom-behavior intervention effective across cultures. *School Psychology International, 35*, 191–205. <http://doi.org/10.1177/0143034312471473>
- Office for Standards in Education. (2013). *The report of Her Majesty's Chief Inspector of Education, Children's Services and Skills*. London: Ofsted.
- Office for Standards in Education. (2014). *Below the radar: low-level disruption in the country's classrooms*. London: Ofsted.
- Oliver, R. M., Wehby, J. H., & Reschly, D. J. (2011). The effects of teachers classroom management practices on disruptive or aggressive student behavior. *The Campbell Systematic Reviews, 44*, 55. <http://doi.org/10.4073/csr.2011.4>
- Pampaka, M., Hutcheson, G., & Williams, J. (2016). Handling missing data: analysis of a challenging data set using multiple imputation. *International Journal of Research & Method in Education, 39*, 19–37. <http://doi.org/10.1080/1743727X.2014.979146>
- Panayiotou, M., & Humphrey, N. (2017). Mental health difficulties and academic attainment: Evidence for gender-specific developmental cascades in middle childhood. *Development and Psychopathology, Early View*, 1–16. <http://doi.org/10.1017/S095457941700102X>
- Pennington, B., & McComas, J. J. (2017). Effects of the good behavior game across classroom contexts. *Journal of Applied Behavior Analysis, 50*, 176–180. <http://doi.org/10.1002/jaba.357>
- Pérez, V., Rodríguez, J., De la Barra, F., & Fernández, A. M. (2005). Efectividad de una estrategia conductual para el manejo de la agresividad en escolares de enseñanza básica. *Psykhe (Santiago), 14*, 55–62. <http://doi.org/10.4067/S0718-22282005000200005>
- Petticrew, M., Tugwell, P., Kristjansson, E., Oliver, S., Ueffing, E., & Welch, V. (2012). Damned if you do, damned if you don't: subgroup analysis and equity. *Journal of Epidemiology and Community Health, 66*, 95–8. <http://doi.org/10.1136/jech.2010.121095>
- Philips Smith, E., Wise, E., Rosen, H., Rosen, A., Childs, S., & McManus, M. (2014). Top-down, bottom-up, and around the Jungle Gym: A social exchange and networks approach to engaging afterschool programs in implementing evidence-based practices. *American Journal of Community Psychology, 53*, 491–502. <http://doi.org/10.1007/s10464-014-9656-0>
- Phillips, D., & Christie, F. (1986). Behaviour management in a secondary school classroom: Playing

- the game. *Maladjustment and Therapeutic Education*, 4, 47–53.
- Reupert, A., & Woodcock, S. (2010). Success and near misses: Pre-service teachers' use, confidence and success in various classroom management strategies. *Teaching and Teacher Education*, 26, 1261–1268. <http://doi.org/10.1016/j.tate.2010.03.003>
- Ruiz-Olivares, R., Pino, M. J., & Herruzo, J. (2010). Reduction of disruptive behaviors using an intervention based on the Good Behavior Game and the Say-Do-Report Correspondence. *Psychology in the Schools*, 47, 1046–1058. <http://doi.org/10.1002/pits.20523>
- Saigh, P. A., & Umar, A. M. (1983). The effects of a good behavior game on the disruptive behavior of Sudanese elementary school students. *Journal of Applied Behavior Analysis*, 16, 339–344.
- Sass, D. A., Seal, A. K., & Martin, N. K. (2011). Predicting teacher retention using stress and support variables. *Journal of Educational Administration*, 49, 200–215. <http://doi.org/10.1108/09578231111116734>
- Skinner, B. F. (1945). The operational analysis of psychological terms. *Psychological Review*, 52, 270–277. <http://doi.org/10.1037/h0062535>
- Slavin, R., & Madden, N. A. (2011). Measures inherent to treatments in program effectiveness reviews. *Journal of Research on Educational Effectiveness*, 4, 370–380. <http://doi.org/10.1080/19345747.2011.558986>
- Smedler, A. C., Hjern, A., Wiklund, S., Anttila, S., & Pettersson, A. (2015). Programs for prevention of externalizing problems in children: Limited evidence for effect beyond 6 months post intervention. *Child and Youth Care Forum*, 44, 251–276. <http://doi.org/10.1007/s10566-014-9281-y>
- Spilt, J. L., Koot, J. M., & Lier, P. A. C. (2013). For whom does it work? Subgroup differences in the effects of a school-based universal prevention program. *Prevention Science*, 14, 479–488. <http://doi.org/10.1007/s11121-012-0329-7>
- Stone, L. L., Otten, R., Engels, R. C. M. E., Vermulst, A. A., & Janssens, J. M. A. M. (2010). Psychometric properties of the parent and teacher versions of the Strengths and Difficulties Questionnaire for 4- to 12-Year-Olds: A Review. *Clinical Child and Family Psychology Review*, 13, 254–274. <http://doi.org/10.1007/s10567-010-0071-2>
- Swiezy, N. B., Matson, J. L., & Box, P. (1993). The Good Behavior Game. *Child & Family Behavior Therapy*, 14, 21–32. [http://doi.org/10.1300/J019v14n03\\_02](http://doi.org/10.1300/J019v14n03_02)
- Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th Editio). Boston, MA: Pearson Education.
- 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. <http://doi.org/10.1016/j.jsp.2010.06.001>
- Tingstrom, D. (1994). The Good Behavior Game: an investigation of teachers' acceptance. *Psychology in the Schools*, 31, 57–65.
- Tingstrom, D. H., Sterling-Turner, H. E., & Wilczynski, S. M. (2006). The good behavior game: 1969-2002. *Behavior Modification*, 30, 225–53. <http://doi.org/10.1177/0145445503261165>
- Tschannen-Moran, M., & Hoy, A. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and Teacher Education*, 17, 783–805.
- Tymms, P. (2004). Effect sizes in multilevel models. In I. Schegen & K. Elliot (Eds.), *But what does it mean? The use of effect sizes in educational research* (pp. 55–65). London: NFER.
- US Department of Health and Human Services. (2002). *Finding the balance: program fidelity and*



*adaptation in substance abuse prevention*. Rockville, MD: US Department of Health and Human Services.

- Varadhan, R., Segal, J. B., Boyd, C. M., Wu, A. W., & Weiss, C. O. (2013). A framework for the analysis of heterogeneity of treatment effect in patient-centered outcomes research. *Journal of Clinical Epidemiology*, *66*, 818–825. <http://doi.org/10.1016/j.jclinepi.2013.02.009>
- Vuijk, P., van Lier, P. A. C., Crijnen, A. A. M., & Huizink, A. C. (2007). Testing sex-specific pathways from peer victimization to anxiety and depression in early adolescents through a randomized intervention trial. *Journal of Affective Disorders*, *100*, 221–226. <http://doi.org/10.1016/j.jad.2006.11.003>
- Warren, S. F., Fey, M. E., & Yoder, P. J. (2007). Differential treatment intensity research: a missing link to creating optimally effective communication interventions. *Mental Retardation and Developmental Disabilities Research Reviews*, *13*, 313–320. <http://doi.org/10.1002/mrdd>
- Weare, K., & Nind, M. (2011). Mental health promotion and problem prevention in schools: what does the evidence say? *Health Promotion International*, *26 Suppl 1*, i29-69. <http://doi.org/10.1093/heapro/dar075>
- Webster, J. B. (1989). Applying behavior management principles with limited resources: Going it alone. *Maladjustment and Therapeutic Education*, *7*, 30–38.
- Weis, R., Osborne, K. J., & Dean, E. L. (2015). Effectiveness of a universal, interdependent group contingency program on children's academic achievement: A countywide evaluation. *Journal of Applied School Psychology*, *31*, 199–218. <http://doi.org/10.1080/15377903.2015.1025322>
- Werthamer-Larsson, L., Kellam, S., & Wheeler, L. (1991). Effect of first-grade classroom environment on shy behavior, aggressive behavior, and concentration problems. *American Journal of Community Psychology*, *19*, 585–602. <http://doi.org/10.1007/BF00937993>
- Whear, R., Thompson-Coon, J., Boddy, K., Ford, T., Racey, D., & Stein, K. (2013). The effect of teacher-led interventions on social and emotional behaviour in primary school children: A systematic review. *British Educational Research Journal*, *39*, 1–38. <http://doi.org/10.1080/01411926.2011.650680>
- Wolf, M. M. (1978). Social validity: the case for subjective measurement. *Journal of Applied Behavior Analysis*, *11*, 203–214.

## Appendix 1: Memorandum of agreement



# Memorandum of Agreement

This Memorandum of Agreement outlines the key conditions for schools entering into partnership with Mentor UK in evaluation of the Good Behaviour Game (GBG). It outlines what schools that participate in the project will receive, and what they will be required to do in return. The aim is to have a completely transparent process so that all parties have a clear understanding of the project and shared expectations.

### Section A: About Your School

We need some key details about your school – please complete the form below:

<b>Name of school</b>	
<b>LAESTAB code</b>	
<b>Address of school</b>	
<b>Postcode of school</b>	
<b>Telephone number of school</b>	
<b>Name of Head Teacher</b>	
<b>Email address of Head Teacher</b>	

### Section A: Your GBG Co-ordinator

It is useful in projects like this to have a nominated 'link' person, who can co-ordinate the project within the school and act as our first point of contact. This GBG co-ordinator could be the head teacher, deputy/assistant head, Key Stage 2 co-ordinator, or a class teacher from Year 3. Please provide details of the nominated link person and the Year 3 teacher(s) for school year 2015/16 below:

# GBG

Good Behaviour Game

<b>Name of GBG co-ordinator</b>	
<b>Email address of GBG co-ordinator</b>	
<b>Primary role within school</b>	
<b>Name(s) of Year 3 teachers 2015/16</b>	
<b>Email address(es) of Year 3 teacher(s) 2015/16</b>	

## Section C: Information about the UK trial

### Aims of the evaluation

The aim of this project is to evaluate the impact of The Good Behaviour Game (GBG), a mixture of comprehensive CPD for teachers combined with a classroom intervention designed to improve children's classroom behaviour in order to improve attainment. The results of the research will contribute to our understanding of what works in raising pupils' attainment. Ultimately, the GBG aims to equip school staff with the skills and materials to better manage classroom behaviour and support children's overall progress.

The evaluation is being conducted by the Manchester Institute of Education at The University of Manchester, and is funded by the Education Endowment Foundation (EEF) and the National Institute for Health Research (NIHR).

### The project

Following baseline data collection in the summer term 2015, schools will be randomly assigned to implement the GBG or continue with their usual practice. Schools randomly assigned to the GBG group begin implementation, starting with training and classroom delivery to Year 3 classes in September 2015. Year 4 teachers in 2016 will be trained either by cascading from the Year 3 teachers using materials provided by the project or

# GBG

Good Behaviour Game

directly by the project's trainers/coaches. The intervention is being delivered by Mentor UK who will also supply all GBG schools with the necessary materials, training and support for classroom delivery of the GBG.

## Structure of the evaluation

A 2-year cluster-randomised trial will be used with randomisation at school level being undertaken by a statistician who is independent of the evaluator. Alongside this a process evaluation will also be undertaken. This means that all schools who decide to participate agree that they can be **randomly assigned** to either (a) **implement the Good Behaviour Game with pupils in Year 3**, or (b) **be a comparison school to continue their usual practice** over a **two-year period** (September 2015 – July 2017).

The trial protocol is available at [www.goodbehaviourgame.info](http://www.goodbehaviourgame.info).

**Random allocation is essential to the evaluation as it is the best way of outlining what effect the GBG has on children's behaviour and attainment. It is important that schools understand and consent to this process.**

## Section D: Key Conditions of Project Participation

In this section we outline the key conditions of project participation. Please read through them carefully.

### All schools

Pay an application fee relevant to the size of their school, this is £750 for single form entry school, £1500 for double form entry school and £2250 for triple form entry school to Mentor UK. This payment is towards the costs of the GBG materials and coaching.

Randomisation – all schools signing this document agree that they can be randomly allocated to either (a) implement the GBG from September 2015 to July 2017 or (b) be a comparison school which continues their usual practice during this period.

Focus – this project focuses on pupils who will be on roll in Year 3 at the start of the 2015/16 school year only.

Compliance with data collection requirements – all schools signing this document understand that they are committing to participation in a research project with certain data collection requirements (see attached flowchart). These are:

# GBG

Good Behaviour Game

1. Annual teacher-pupil surveys to be conducted in the summer term (typically in May-July) of the school years 2014/15, 2015/16, 2016/17, 2017/18 and 2018/19.
2. Annual staff survey in the summer terms (May-July) of school years 2014/15, 2015/16 and 2016/17.
3. Pupil survey and reading test in summer term (May-July) of schools years 2016/17, 2017/18 and 2018/19.

Schools must complete at least **90% of surveys** in May-July 2014/15 in order to be eligible for randomisation.

## Schools randomly assigned to the GBG group:

For schools randomly allocated to the GBG group only, a commitment to implement the programme throughout the school years 2015/16 and 2016/17 is required. This includes GBG schools undertaking to ensure that all of their Year 3 teachers are available for two days of training on the GBG in September 2015 and for one day's booster training later in the school year.

GBG schools will not have to make any additional payment to Mentor UK during the two years of the GBG beyond their payment made as a part of the application process.

## Schools randomly assigned to the comparison group:

Schools randomly allocated to the comparison group will continue practice as usual during the school years 2015/16 and 2016/17, and their application fee will be returned to them.

## Section E: What Participating Schools Will Receive

This section outlines what each participating school will receive as part of the project.

### All participating schools will receive:

Generic feedback from our pupil-teacher surveys following each annual wave of data collection.

### In addition, schools randomly allocated to the GBG group only will receive:

- GBG training for teachers of children in Years 3 in school year 2015/16
  1. 2 days initial training during September 2015
  2. 1 day follow-up training
- The GBG materials
- The intervention is being delivered by Mentor UK who will also supply all GBG schools with the necessary materials, training and coaching for classroom delivery of the GBG.
- GBG training for teachers of children in Years 4 in school year 2016/17. Training will be either by cascading from the Year 3 teachers using materials provided by the project or directly by the project's trainers/coaches.

### In addition, schools randomly allocated to the comparison group only will receive:

A payment of £1500 will also be made to comparison schools for their participation in the GBG trial and the data collection associated with this - £1000 at the beginning of the trial and £250 when all data has been submitted at the end of the second year of the trial, with a further £250 when all data has been submitted at the end of the fourth year of the trial. The above amounts are for two form entry schools and will be pro rata for one and three form entry schools.

### Use of data

Pupils' test responses and any other pupil data will be treated with the strictest confidence. The responses will be collected online and/or on paper. The website that houses the survey will be completely secure and password protected. All survey data will be stored on a secure, password protected computer to which only senior members of the research team have access. Named data will be matched with the National Pupil Database and shared with Mentor UK, the EEF and the NIHR. No individual school or pupil will be identified in any report arising from the research.

## Section F: Commitment to Participate

### Application requirements

To complete their application for the GBG, schools need to:

- Send the completed Memorandum of Agreement to Mentor UK (details below) signed by the school's Head Teacher.
- Make payment relevant to the size of their school, this is £750 for single form entry school, £1500 for double form entry school and £2250 for triple form entry school.
- Complete the collection of baseline data, May-July 2014/15, prior to randomisation in the summer term. Schools must complete at least 90% of surveys in order to be eligible for randomisation.

### Commitment to participate

We confirm that we have read and understood all of the above and are happy to confirm our participation in trial of the Good Behaviour Game as per the details specified, on behalf of

School: \_\_\_\_\_

\_\_\_\_\_  
Headteacher (signature)                      \_\_\_\_\_                      \_\_\_\_\_  
Print name    Date

\_\_\_\_\_  
Chair of Board of Governors (sig.)                      \_\_\_\_\_                      \_\_\_\_\_  
Print name    Date

\_\_\_\_\_  
                      \_\_\_\_\_  
Simon Claridge

\_\_\_\_\_  
Director of Programmes, Mentor UK                      \_\_\_\_\_                      \_\_\_\_\_  
Print name    Date

Please sign two copies, retaining one and returning the second copy to:

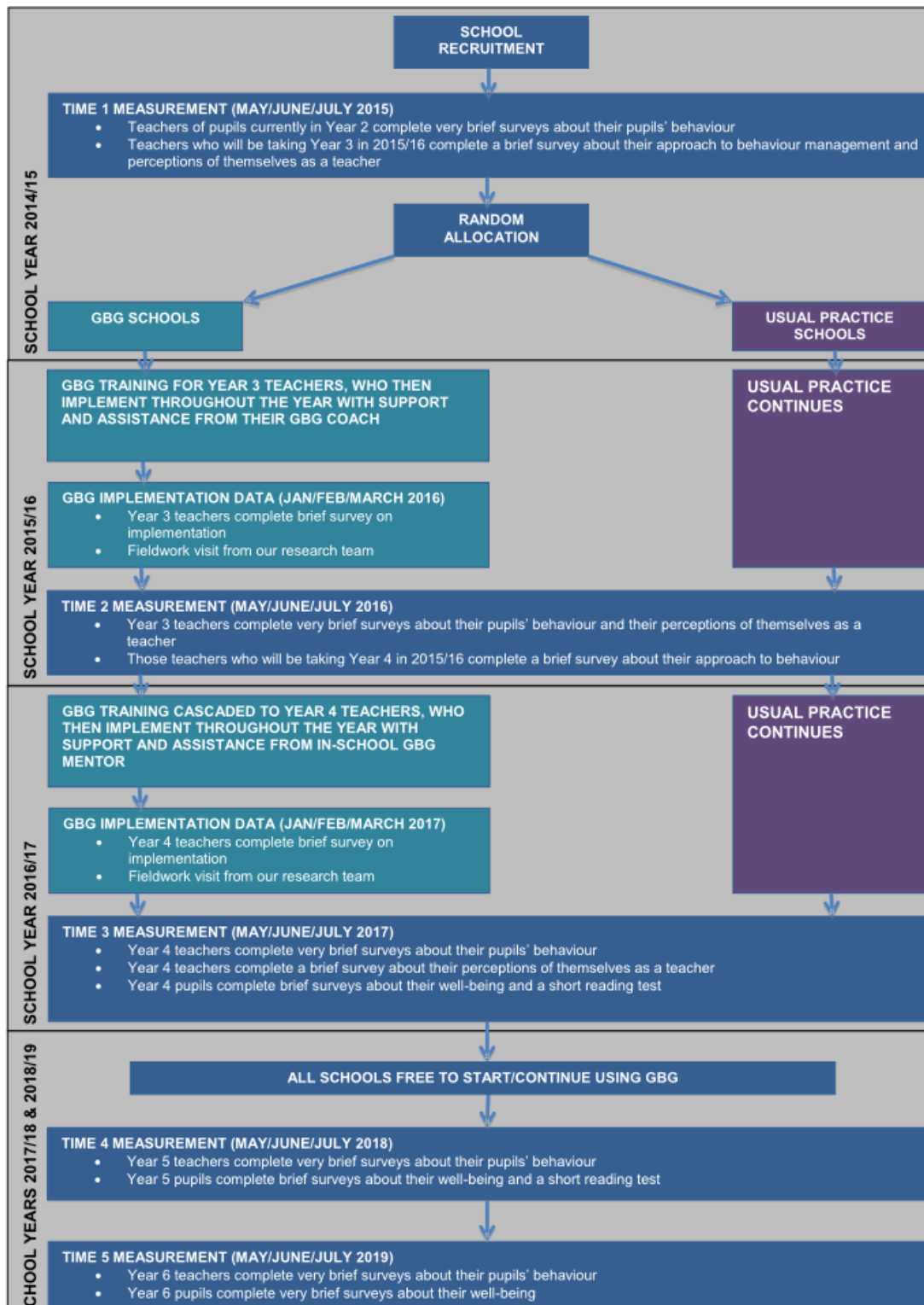
Amanda Hood  
Mentor UK  
CAN-Mezzanine  
49-51 East Road  
London, N1 6AH



# GBG

Good Behaviour Game

## Good Behaviour Game Trial: Basic Study Flowchart





## Appendix 2: Information sheet for parents

# The Good Behaviour Game

## INFORMATION SHEET FOR PARENTS

Your child's school is involved in a project about the Good Behaviour Game. The Good Behaviour Game is a way to help children to concentrate on their school work and improve their behaviour. It has been shown to be very helpful in other countries in the world. We want to find out if it can help children in England too. The project is funded by The Education Endowment Foundation and The National Institute for Health Research.

We are writing to you because your child's school is involved in the project. We will ask your child's teacher to complete a survey about your child's behaviour once a year starting summer (May-July) 2015. From the summer of 2017 onwards we will also ask your child to complete a brief annual survey about their wellbeing (see below for more details). Our surveys will conclude in summer 2019.

Please take time to read the following information carefully and decide whether or not your child would like to take part.

If you would like any more information or have any questions about the research project, please telephone Dr. Alexandra Barlow on 0161 275 3504 or email her at [alexandra.barlow@manchester.ac.uk](mailto:alexandra.barlow@manchester.ac.uk).

### **Who will conduct the research?**

The research will be conducted by Professor Neil Humphrey and his research team at the Manchester Institute of Education, The University of Manchester, Oxford Road, Manchester M13 9PL.

### **Title of the research**

"The Good Behaviour Game"

### **What is the aim of the research?**

Our main aim is to examine the impact of the Good Behaviour game on reading and behaviour.

**Where will the research be conducted?**

Primary schools in Greater Manchester, West Yorkshire, South Yorkshire and East Midlands.

**What is the duration of the research?**

The project itself runs from September 2014 until March 2020. The schools that implement the Good Behaviour Game (see below) will do so from September 2015 to July 2017.

**Why have I been chosen?**

We are writing to you because your child's school is taking part in the Good Behaviour Game project. Schools will be randomly chosen to (a) implement the Good Behaviour Game over a two year period (Good Behaviour Game schools), or (b) continue as normal (comparison schools). We will be collecting data in both Good Behaviour Game and comparison schools. After two years, all schools will be free to decide whether they wish to start/continue using the Good Behaviour Game.

**What would my child be asked to do if he/she took part?**

Your child's class teacher will be asked to complete a brief online survey about your child's behaviour. These surveys will be completed annually – in May/-July 2015, 2016, 2017, 2018 and 2019.

Your child will be asked to complete both a short reading assessment and a short survey about wellbeing at the end of the main trial in summer (May-July) 2017, and again in May-July 2018 and 2019. The survey will take approximately 20 minutes to complete and the reading assessment will take approximately 30 minutes to complete.

If you agree, you will be saying that your child can take the tests and fill in the questionnaires. You will also be saying that his/her teacher can complete surveys about him/her.

In consenting to your child's participation, you are also giving permission that for the purpose of the study, information provided will be linked with the National Pupil Database (held by the Department for Education), other official records, and shared with the Department for Education, Education Endowment Foundation (EEF), EEF's data contractor FFT Education, and in an anonymised form to the UK Data Archive.

### **What happens to the data collected?**

The data will be downloaded from our secure online survey site so that it can be analysed by our research team at the University of Manchester. We will write a report based on our analyses for our funders, the Education Endowment Foundation and the National Institute for Health Research. It is also likely that we will write articles for academic journals based on what we find out in the project. The data may also be used as part of a doctoral thesis. Finally, it is possible that we will write a book about the research. Your child's name will not be used in any of the reports that we write.

### **How is confidentiality maintained?**

All data provided will be treated as confidential and will be completely anonymous. Identifying information (e.g. your child's name) will only be used in order to match responses about the same individual from different respondents (e.g. teacher and pupil surveys) and across different times (e.g. May-July 2015, 2016, and 2017). After this matching process is complete, all identifying information will be destroyed.

The website that houses these surveys will be completely secure and password protected. All survey data will be stored on a secure, password protected computer to which only senior members of the research team have access.

### **What happens if I do not want my child to take part or I change my mind later?**

It is up to you if you want your child to take part in the data collection.

If you decide your child and his/her teacher can take part in the data collection you do not need to do anything – your child's school will be sent further details about when and how to complete the survey in the near future.

If you decide not to take part then you need to either complete the opt-out consent form enclosed and return it to our research team or contact Dr. Alexandra Barlow by telephone or email (details below).

If you decide to take part and then change your mind, you are free to withdraw without needing to give a reason by contacting Dr. Alexandra Barlow by telephone or email (details below). We will send annual reminders about the study, but you can opt your child out at any time up until the end of the study, in summer 2017. If you do this please rest assured that we will destroy any data collected about your child as part of the study.

### **Will I be paid for participating in the research?**

We are not able to offer any payment or incentive for participating in this study.

### **Disclosure and Barring Service (DBS) Check**

Every member of our research team has undergone a Disclosure and Barring Service (formerly 'Criminal Records Bureau') check at the Enhanced Disclosure level. This means that they have permission to work with and do research with children.

### **Contact for further information**

Dr. Alexandra Barlow

Educational Support and Inclusion

School of Education

University of Manchester

Oxford Road

Manchester

M13 9PL

Tel: 0161 275 3504

Email: [alexandra.barlow@manchester.ac.uk](mailto:alexandra.barlow@manchester.ac.uk)

Also, please see our website for further details about the Good Behaviour Game and background, the project design and project team.

The website can be found at: <http://www.goodbehaviourgame.info>

### **What if something goes wrong?**

If your child or your child's teacher completing the survey makes you worry about your child's wellbeing then you should contact the school in the first instance and ask to speak to his/her teacher.

You can also get independent support and advice from a charity called Young Minds. Their parent helpline number is 0808 802 5544.

### **What if I want to complain?**

If you have any concerns or wish to complain, you should contact the researcher Alexandra Barlow in the first instance (contact details above).

If you remain dissatisfied, or if the research team is unable to address the issues you raise you should contact the Head of School, Prof Tim Allott (School of Environment, Education and Development), at [Tim.Allott@manchester.ac.uk](mailto:Tim.Allott@manchester.ac.uk) or on 0161 275 3662.

If there are any issues regarding this research that you would prefer not to discuss with members of the research team or Head of School, please contact the Research Practice and Governance Co-ordinator by either writing to 'The Research Practice and Governance Co-ordinator, Research Office, Christie Building, The University of Manchester, Oxford Road, Manchester M13 9PL', by emailing: [Research-Governance@manchester.ac.uk](mailto:Research-Governance@manchester.ac.uk), or by telephoning 0161 275 7583 or 275 8093

## Appendix 3: Parent consent form



### The Good Behaviour Game

## CONSENT FORM FOR PARENTS

An information sheet is attached to this form. Please read it carefully before making a decision about taking part.

If you are willing to let your child take part and for his/her teacher to give information about him/her then you do not need to do anything at the moment.

If you decide not to let your child take part, then you need to complete the opt-out consent form below and use the freepost code below to return it to us:

FREEPOST RLYU-KAAB-AXRC

Dr. Alexandra Barlow,

Manchester Institute of Education

The University of Manchester,

Ellen Wilkinson Building

Oxford Road,

Manchester,

M13 9PL.

Alternatively, Dr. Barlow can be contacted by telephone on 0161 275 3504 or email at [alexandra.barlow@manchester.ac.uk](mailto:alexandra.barlow@manchester.ac.uk). If you do not want your child to participate please let us know by Friday 2<sup>nd</sup> October 2015.

Finally, please also remember that if you do decide he/she can take part, you are free to change your mind at any point in the study.

-----

I do not wish my child to participate in the Good Behaviour Game project. My details are as follows:

My name	
My child's name	
Name of my child's school	

Signed: \_\_\_\_\_

Date: \_\_\_\_\_

## Appendix 4: Confirmatory factor analyses for teacher usual practice surveys

Table A1: Eigenvalues from Parallel Analysis for each subscale

Subscale	EV	50 <sup>th</sup> percentile EV	95 <sup>th</sup> percentile EV
<b>GBG-GB</b>			
<b>Factor 1</b>	<b>3.51</b>	1.50	1.60
<b>Factor 2</b>	<b>1.76</b>	1.40	1.47
<b>Factor 3</b>	1.33	1.32	1.38
<b>GBG-RS</b>			
<b>Factor 1</b>	<b>2.73</b>	1.33	1.43
<b>Factor 2</b>	1.10	1.23	1.30
<b>GBG-DB</b>			
<b>Factor 1</b>	<b>4.40</b>	1.59	1.69
<b>Factor 2</b>	<b>2.16</b>	1.48	1.56
<b>Factor 3</b>	1.44	1.41	1.47

Note. EV = eigenvalues.

In bold are factors with eigenvalues larger than the 50<sup>th</sup> and 95<sup>th</sup> percentile ones.

Based on findings from parallel analysis, a 2-factor ESEM was conducted for the 22-item general behaviour management and is summarised in the Table below. Items 12 and 16 were removed from the analysis as they failed to substantially load onto any of the two factors. While many of the items were shown to have high factor loadings, they failed to reach statistical significance, which may be due to the small sample size (65 parameters for n=245). The sample size for this analysis is considered to be lower than the minimum suggested 5:1 criterion (i.e. 5 participants per parameter) thus requiring caution when interpreting the results. Therefore, a one-factor ESEM with less parameters was also explored (with 6:1), results of which point to a good fit. Items 3,9,11, and 13 were not retained due to poor factor loadings.



**Table A2: 2-Factor ESEM for General behaviour management approaches subscale**

		Factor 1 <sup>a</sup>	Factor 2 <sup>b</sup>
1.	I establish and maintain a set of classroom rules	-.001	<b>.447***</b>
2.	My pupils help to establish the rules of the classroom	-.139	<b>.482***</b>
3.	I use behaviour contracts	<b>.391</b>	-.220
4.	I communicate clear expectations about rules and pupils' responsibilities e.g. through posters	<b>.441</b>	.214
5.	I give pupils positions of responsibility	.240	<b>.367</b>
6.	I alter the seating plan in my classroom as part of my behaviour management strategy	.293	<b>.451</b>
7.	I alter the curriculum to match pupils' interests and needs	.037	<b>.507***</b>
8.	I promote good behaviour through PSHE lessons	<b>.490**</b>	.187
9.	I use Circle Time to promote and help understanding of good behaviour	<b>.568**</b>	.033
10.	I incorporate teaching of appropriate behaviours in lessons e.g. prosocial behaviours such as teamwork	<b>.507*</b>	.314
11.	I attend behaviour management courses/CPD	<b>.371</b>	-.051
13.	I use buddying/peer mentoring techniques	<b>.443**</b>	-.002
14.	I use targeted behaviour management strategies for specific pupils	.322	<b>.405</b>
15.	I follow my school's behaviour policy	.400	<b>.623</b>
17.	I use signals e.g. clapping	.199	<b>.364</b>
18.	I use verbal redirection to engage pupils	.304	<b>.358</b>
19.	I reinforce our whole school behaviour policy/ethos/values	-.017	<b>.826***</b>
20.	I focus on good behaviour – “catch them doing the right thing”	-.017	<b>.826***</b>
21.	I observe and monitor pupils' behaviour in the classroom	.335	<b>.531</b>
22.	I respond to disruptive behaviour promptly	.376	<b>.589</b>

Note. a = KR-20 = .55; b = KR-20 = .35

Number of imputed datasets = 50.

$\chi^2_{\text{mean}}(151) = 163.641$  ( $SD = 11.128$ ); Mean RMSEA = .017 ( $SD = .007$ ); Mean CFI = 985 ( $SD = .013$ ); Mean TLI = .982 ( $SD = .016$ ).

**Table A3: 1-Factor ESEM for General behaviour management approaches subscale**

	Factor 1 <sup>a</sup>
1. I establish and maintain a set of classroom rules	<b>.477***</b>
2. My pupils help to establish the rules of the classroom	<b>.443***</b>
4. I communicate clear expectations about rules and pupils' responsibilities e.g. through posters	<b>.501***</b>
5. I give pupils positions of responsibility	<b>.532***</b>
6. I alter the seating plan in my classroom as part of my behaviour management strategy	<b>.638***</b>
7. I alter the curriculum to match pupils' interests and needs	<b>.471***</b>
8. I promote good behaviour through PSHE lessons	<b>.458***</b>
10. I incorporate teaching of appropriate behaviours in lessons e.g. prosocial behaviours such as teamwork	<b>.638***</b>
12. I use an anti-bullying policy	<b>.517***</b>
14. I use targeted behaviour management strategies for specific pupils	<b>.620***</b>
15. I follow my school's behaviour policy	<b>.981***</b>
16. I use the "silent and still" approach – stopping and waiting for pupils to respond	<b>.346***</b>
17. I use signals e.g. clapping	<b>.536***</b>
18. I use verbal redirection to engage pupils	<b>.541***</b>
19. I reinforce our whole school behaviour policy/ethos/values	<b>.935***</b>
20. I focus on good behaviour – "catch them doing the right thing"	<b>.935***</b>
21. I observe and monitor pupils' behaviour in the classroom	<b>.806***</b>
22. I respond to disruptive behaviour promptly	<b>.895***</b>

Note. a = KR-20 = .51

Number of imputed datasets = 50.

$\chi^2$ mean(135) = 152.162 (SD = 3.00); Mean RMSEA = .023 (SD = .002) ; Mean CFI = .978 (SD = .003); Mean TLI = .975 (SD = .003).

The 1-factor reward systems subscale was shown to have an acceptable structure. The majority of the items had substantial and statistically significant loadings, except for items 1 and 9, which were removed from the scale. The internal consistency of the scale was also found to be acceptable.

**Table A4: 1-Factor EFA for Rewards Systems subscale**

	<b>Factor 1</b>
2. I use an educational reward system e.g. free time, time on the computer	<b>0.642<sup>***</sup></b>
3. I use prizes as rewards for good behaviour	<b>0.476<sup>***</sup></b>
4. I use individual rewards	<b>0.645<sup>***</sup></b>
5. I use group rewards	<b>0.699<sup>***</sup></b>
6. I use whole class rewards	<b>0.707<sup>***</sup></b>
7. I use special privileges	<b>0.534<sup>***</sup></b>
8. I send notes/call/text parents about good behaviour	<b>0.381<sup>***</sup></b>
10. I/we hold assemblies in which good behaviour is recognised/rewarded, e.g. giving of certificates	<b>0.319<sup>***</sup></b>

Note. <sup>\*\*\*</sup>  $p < .001$ .

$\alpha = .71$ , Raykov  $\omega = .72$

A similar structure with similar factor loadings was observed in multiple imputation ESEM.  $\chi^2(20) = 47.969$ ,  $p < .001$ ; RMSEA = .076 (.048, .103),  $p > .05$ ; CFI = .936; TLI = .911.

Similarly, all items (except for 1 and 2) were shown to have substantial and statistically significant factor loading on the 2-factor disruptive behaviour management. Item 17 was not retained because of empty cells in the bivariate table. This could be explained by the fact that there were data on only two of the four response categories. The rotated structure is shown in Table A5 below.

**Table A5: 2-Factor EFA for Managing Disruptive and Inappropriate Behaviour subscale**

		Factor 1 <sup>a</sup>	Factor 2 <sup>b</sup>
3.	I use vocal warnings e.g. raising/lowering voice, shouting	<b>0.737***</b>	-0.005
4.	I use body language e.g. frowning, physical proximity	<b>0.830***</b>	-0.143
5.	I remove privileges	<b>0.547***</b>	0.198
6.	I use threats e.g. removal of rewards	<b>0.708***</b>	0.019
7.	I use a warning/strike system	0.301***	<b>0.338**</b>
8.	I move pupils who are misbehaving to a different area of the classroom/make them stand up/send them out of the classroom	0.270**	<b>0.502***</b>
9.	I single out a child/group of children for misbehaviour	<b>0.522***</b>	0.124
10.	I use a restorative justice system	0.119	<b>0.437***</b>
11.	I use break-time supervision	0.115	<b>0.540***</b>
12.	I use detention	0.042	<b>0.330**</b>
13.	I use a behaviour report card	-0.063	<b>0.516***</b>
14.	I use a behaviour support base in the school	-0.152*	<b>0.581***</b>
15.	I contact pupils' parents/carers	-0.021	<b>0.596***</b>
16.	I refer pupils to the Head Teacher/other professionals	0.003	<b>0.823***</b>

Note. \*  $p < .05$  \*\*  $p < .01$  \*\*\*  $p < .001$ .

$X^2(64) = 120.882$ ,  $p < .001$ ; RMSEA = .060 (.044, .077),  $p > .05$ ; CFI = .924; TLI = .891.

A similar structure with similar factor loadings was observed in multiple imputation ESEM.

a =  $\alpha = .64$  ; Raykov  $\omega = .61$

b =  $\alpha = .53$  ; Raykov  $\omega = .72$

# Appendix 5: GBG observation schedule

<p><b>1.</b></p> <p>Date (dd/mm/yyyy)</p> <p>School Code</p> <p>Teacher Code</p> <p>Start Time (hh:mm)</p> <p>End Time (hh:mm)</p> <p>Observer Name</p> <p>Location</p> <p><b>3. Pre-Game</b></p> <p>Lesson</p> <p>Task/Activity</p> <p>Independent/Group/Pair Work</p> <p>Voice Level</p> <p>Type of Timer</p> <p>Game Length (mins)</p> <p>N Pupils per Team (Range)</p> <p>N Teams in the Class</p> <table border="1" style="width:100%; border-collapse: collapse;"> <thead> <tr> <th>Descriptor</th> <th>Procedural Fidelity</th> <th>Quality</th> </tr> </thead> <tbody> <tr> <td><b>Activity</b></td> <td></td> <td></td> </tr> <tr> <td>Teacher explains the task/activity</td> <td>Y N</td> <td>1 2</td> </tr> <tr> <td>Teacher checks understanding of the task/activity</td> <td>Y N</td> <td>1 2</td> </tr> <tr> <td>Teacher reminds pupils that they cannot ask for help</td> <td>Y N</td> <td>1 2</td> </tr> <tr> <td><b>Teams</b></td> <td></td> <td></td> </tr> <tr> <td>Pupils are in teams of between 3 and 7 (except for special circumstances, e.g. team of 1)</td> <td>Y N</td> <td>1 2</td> </tr> <tr> <td>Pupils are in clear teams</td> <td>Y N</td> <td>1 2</td> </tr> <tr> <td>Teams are gender balanced</td> <td>Y N</td> <td>1 2</td> </tr> <tr> <td><b>Rules</b></td> <td></td> <td></td> </tr> <tr> <td>Rules appropriately verbally reviewed with class</td> <td>Y N</td> <td>1 2</td> </tr> <tr> <td>Exemplars modelled/described by teacher and/or students</td> <td>Y N</td> <td>1 2</td> </tr> <tr> <td>Infractions modelled/described by teacher</td> <td>Y N</td> <td>1 2</td> </tr> <tr> <td>Infractions only described by students</td> <td>Y N</td> <td>1 2</td> </tr> <tr> <td>Voice level given by teacher</td> <td>Y N</td> <td>1 2</td> </tr> <tr> <td><b>Game Specifics</b></td> <td></td> <td></td> </tr> <tr> <td>Teacher states when the game begins</td> <td>Y N</td> <td></td> </tr> <tr> <td>Teacher states how long the game will be played for</td> <td>Y N</td> <td></td> </tr> <tr> <td>Teacher sets timer</td> <td>Y N</td> <td></td> </tr> <tr> <td>Teacher states that they will monitor infractions</td> <td>Y N</td> <td></td> </tr> <tr> <td>Teacher states that 4 infractions are permitted</td> <td>Y N</td> <td></td> </tr> <tr> <td>Teacher reminds pupils that they are not competing against each other</td> <td>Y N</td> <td></td> </tr> <tr> <td>Notes</td> <td colspan="2">Interpretation</td> </tr> <tr> <td>Adaptations</td> <td>Notes</td> <td>Interpretation</td> </tr> <tr> <td>Teams</td> <td></td> <td></td> </tr> <tr> <td>Rules</td> <td></td> <td></td> </tr> <tr> <td>Activity</td> <td></td> <td></td> </tr> <tr> <td>Game specifics</td> <td></td> <td></td> </tr> <tr> <td>How do pupils respond to the announcement of the game?</td> <td>N/A</td> <td>0 1 2</td> </tr> <tr> <td>How attentive are pupils to the teacher's instructions and examples regarding the game?</td> <td>N/A</td> <td>0 1 2</td> </tr> <tr> <td>How enthusiastic/willing to participate are pupils when discussing the rules?</td> <td>N/A</td> <td>0 1 2</td> </tr> <tr> <td>Notes</td> <td colspan="2">Interpretation</td> </tr> </tbody> </table>	Descriptor	Procedural Fidelity	Quality	<b>Activity</b>			Teacher explains the task/activity	Y N	1 2	Teacher checks understanding of the task/activity	Y N	1 2	Teacher reminds pupils that they cannot ask for help	Y N	1 2	<b>Teams</b>			Pupils are in teams of between 3 and 7 (except for special circumstances, e.g. team of 1)	Y N	1 2	Pupils are in clear teams	Y N	1 2	Teams are gender balanced	Y N	1 2	<b>Rules</b>			Rules appropriately verbally reviewed with class	Y N	1 2	Exemplars modelled/described by teacher and/or students	Y N	1 2	Infractions modelled/described by teacher	Y N	1 2	Infractions only described by students	Y N	1 2	Voice level given by teacher	Y N	1 2	<b>Game Specifics</b>			Teacher states when the game begins	Y N		Teacher states how long the game will be played for	Y N		Teacher sets timer	Y N		Teacher states that they will monitor infractions	Y N		Teacher states that 4 infractions are permitted	Y N		Teacher reminds pupils that they are not competing against each other	Y N		Notes	Interpretation		Adaptations	Notes	Interpretation	Teams			Rules			Activity			Game specifics			How do pupils respond to the announcement of the game?	N/A	0 1 2	How attentive are pupils to the teacher's instructions and examples regarding the game?	N/A	0 1 2	How enthusiastic/willing to participate are pupils when discussing the rules?	N/A	0 1 2	Notes	Interpretation		<p><b>2. Classroom</b></p> <p>Number of Children</p> <p>Number of Absences and Withdrawals (note reasons)</p> <p>Number of Staffs Present (other than teacher, e.g. teaching assistants). Note their roles (if known) in the space below</p> <p>Notes</p> <p><b>4. During Game</b></p> <table border="1" style="width:100%; border-collapse: collapse;"> <thead> <tr> <th>Descriptor</th> <th>Procedural Fidelity</th> <th>Quality</th> </tr> </thead> <tbody> <tr> <td><b>Check, Comment, Redirect</b></td> <td></td> <td></td> </tr> <tr> <td>Teacher records majority of infractions on scoreboard</td> <td>Y N</td> <td>1 2</td> </tr> <tr> <td>Teacher identifies majority rule broken</td> <td>Y N</td> <td>1 2</td> </tr> <tr> <td>Teacher discreetly indicates rules broken to specific pupil most of the time</td> <td>Y N</td> <td>1 2</td> </tr> <tr> <td>Teacher frequently identifies rule breaking team</td> <td>Y N</td> <td>1 2</td> </tr> <tr> <td>Rest of team frequently praised for adhering to rules</td> <td>Y N</td> <td>1 2</td> </tr> <tr> <td>Other teams frequently praised for adhering to rules</td> <td>Y N</td> <td>1 2</td> </tr> <tr> <td>Teacher does not punish pupils/teams for infractions</td> <td>Y N</td> <td></td> </tr> <tr> <td><b>Game Management</b></td> <td></td> <td></td> </tr> <tr> <td>Teacher monitors behaviour</td> <td>Y N</td> <td>1 2</td> </tr> <tr> <td>Teacher does not interact with pupils</td> <td>Y N</td> <td></td> </tr> <tr> <td>Teacher adheres to time limit</td> <td>Y N</td> <td></td> </tr> <tr> <td>Teacher announces the end of the game</td> <td>Y N</td> <td></td> </tr> <tr> <td>Notes</td> <td colspan="2">Interpretation</td> </tr> <tr> <td>Adaptations</td> <td>Notes</td> <td>Interpretation</td> </tr> <tr> <td>Check, comment, redirect</td> <td></td> <td></td> </tr> <tr> <td>Game management</td> <td></td> <td></td> </tr> <tr> <td><b>Team</b></td> <td>1</td> <td>2</td> <td>3</td> <td>4</td> <td>5</td> </tr> <tr> <td>Rule 1 We will work quietly</td> <td></td> <td></td> <td></td> <td></td> <td></td> </tr> <tr> <td>Rule 2 We will be polite to others</td> <td></td> <td></td> <td></td> <td></td> <td></td> </tr> <tr> <td>Rule 3 We will get out of our seats with permission</td> <td></td> <td></td> <td></td> <td></td> <td></td> </tr> <tr> <td>Rule 4 We will follow directions</td> <td></td> <td></td> <td></td> <td></td> <td></td> </tr> <tr> <td>TOTAL Infractions</td> <td></td> <td></td> <td></td> <td></td> <td></td> </tr> <tr> <td>Generally, do rule breaking pupils correct their behaviour following an infraction?</td> <td>N/A</td> <td>0 1 2</td> </tr> <tr> <td>Generally, how well do pupils respond to a member of their team getting a check?</td> <td>N/A</td> <td>0 1 2</td> </tr> <tr> <td>Are there any externalising responses from pupils when they receive a check?</td> <td></td> <td></td> </tr> <tr> <td>Notes</td> <td colspan="2">Interpretation</td> </tr> </tbody> </table>	Descriptor	Procedural Fidelity	Quality	<b>Check, Comment, Redirect</b>			Teacher records majority of infractions on scoreboard	Y N	1 2	Teacher identifies majority rule broken	Y N	1 2	Teacher discreetly indicates rules broken to specific pupil most of the time	Y N	1 2	Teacher frequently identifies rule breaking team	Y N	1 2	Rest of team frequently praised for adhering to rules	Y N	1 2	Other teams frequently praised for adhering to rules	Y N	1 2	Teacher does not punish pupils/teams for infractions	Y N		<b>Game Management</b>			Teacher monitors behaviour	Y N	1 2	Teacher does not interact with pupils	Y N		Teacher adheres to time limit	Y N		Teacher announces the end of the game	Y N		Notes	Interpretation		Adaptations	Notes	Interpretation	Check, comment, redirect			Game management			<b>Team</b>	1	2	3	4	5	Rule 1 We will work quietly						Rule 2 We will be polite to others						Rule 3 We will get out of our seats with permission						Rule 4 We will follow directions						TOTAL Infractions						Generally, do rule breaking pupils correct their behaviour following an infraction?	N/A	0 1 2	Generally, how well do pupils respond to a member of their team getting a check?	N/A	0 1 2	Are there any externalising responses from pupils when they receive a check?			Notes	Interpretation		<p><b>Physical Artifacts</b></p> <table border="1" style="width:100%; border-collapse: collapse;"> <thead> <tr> <th></th> <th>PIA</th> <th>P</th> <th>No</th> </tr> </thead> <tbody> <tr> <td>Rules Poster</td> <td></td> <td></td> <td></td> </tr> <tr> <td>Voice Levels Poster</td> <td></td> <td></td> <td></td> </tr> <tr> <td>Team Assignment Wall-Chart</td> <td></td> <td></td> <td></td> </tr> <tr> <td>Scoreboard/Poster</td> <td></td> <td></td> <td></td> </tr> <tr> <td>Rules on Desk</td> <td></td> <td></td> <td></td> </tr> <tr> <td>GBG Booklets</td> <td></td> <td></td> <td></td> </tr> <tr> <td>Timer</td> <td></td> <td></td> <td></td> </tr> <tr> <td>Stamps/Stickers for Booklets</td> <td></td> <td></td> <td></td> </tr> <tr> <td>Reinforcers</td> <td></td> <td></td> <td></td> </tr> <tr> <td>Notes</td> <td colspan="3">Interpretation</td> </tr> </tbody> </table> <p><b>5. Post-Game</b></p> <table border="1" style="width:100%; border-collapse: collapse;"> <thead> <tr> <th>Descriptor</th> <th>Procedural Fidelity</th> <th>Quality</th> </tr> </thead> <tbody> <tr> <td><b>Game Management</b></td> <td></td> <td></td> </tr> <tr> <td>Teacher repeats 4 checks or less criterion</td> <td>Y N</td> <td>1 2</td> </tr> <tr> <td>Teacher announces winning teams only</td> <td>Y N</td> <td>1 2</td> </tr> <tr> <td><b>Reinforcers</b></td> <td></td> <td></td> </tr> <tr> <td>Members of winning team receive stamp (or marker etc)</td> <td>Y N</td> <td>1 2</td> </tr> <tr> <td>Star placed on wall-chart</td> <td>Y N</td> <td>1 2</td> </tr> <tr> <td>Notes</td> <td colspan="2">Interpretation</td> </tr> <tr> <td>Adaptations</td> <td>Notes</td> <td>Interpretation</td> </tr> <tr> <td>Game Management</td> <td></td> <td></td> </tr> <tr> <td>Reinforcers</td> <td></td> <td></td> </tr> <tr> <td>Generally, what is the level of interest and attentiveness to the reinforcers?</td> <td>N/A</td> <td>0 1 2</td> </tr> <tr> <td>Generally, how do team leaders respond to sticking items on the board?</td> <td>N/A</td> <td>0 1 2</td> </tr> <tr> <td>Generally, how do pupils respond if they do not win the game? (note behaviours)</td> <td>N/A</td> <td>0 1 2</td> </tr> <tr> <td>Notes</td> <td colspan="2">Interpretation</td> </tr> <tr> <td><b>Reinforcers</b></td> <td></td> <td></td> </tr> <tr> <td>Type</td> <td>Tangible</td> <td>Intangible</td> <td>Token</td> </tr> <tr> <td>Praise(s) Given</td> <td></td> <td></td> <td></td> </tr> <tr> <td>Delay of Gratification</td> <td>Immediate</td> <td>Delayed</td> <td></td> </tr> <tr> <td>Notes</td> <td colspan="3">Interpretation</td> </tr> </tbody> </table> <p><b>6. Overall - Teacher</b></p> <table border="1" style="width:100%; border-collapse: collapse;"> <tbody> <tr> <td>Interest and enthusiasm</td> <td>0</td> <td>1</td> <td>2</td> </tr> <tr> <td>Clarity of expression</td> <td>0</td> <td>1</td> <td>2</td> </tr> <tr> <td>Preparedness</td> <td>0</td> <td>1</td> <td>2</td> </tr> <tr> <td>Consistency of behaviour</td> <td>0</td> <td>1</td> <td>2</td> </tr> <tr> <td>Engagement of pupils</td> <td>0</td> <td>1</td> <td>2</td> </tr> <tr> <td>Notes</td> <td colspan="3">Interpretation</td> </tr> </tbody> </table>		PIA	P	No	Rules Poster				Voice Levels Poster				Team Assignment Wall-Chart				Scoreboard/Poster				Rules on Desk				GBG Booklets				Timer				Stamps/Stickers for Booklets				Reinforcers				Notes	Interpretation			Descriptor	Procedural Fidelity	Quality	<b>Game Management</b>			Teacher repeats 4 checks or less criterion	Y N	1 2	Teacher announces winning teams only	Y N	1 2	<b>Reinforcers</b>			Members of winning team receive stamp (or marker etc)	Y N	1 2	Star placed on wall-chart	Y N	1 2	Notes	Interpretation		Adaptations	Notes	Interpretation	Game Management			Reinforcers			Generally, what is the level of interest and attentiveness to the reinforcers?	N/A	0 1 2	Generally, how do team leaders respond to sticking items on the board?	N/A	0 1 2	Generally, how do pupils respond if they do not win the game? (note behaviours)	N/A	0 1 2	Notes	Interpretation		<b>Reinforcers</b>			Type	Tangible	Intangible	Token	Praise(s) Given				Delay of Gratification	Immediate	Delayed		Notes	Interpretation			Interest and enthusiasm	0	1	2	Clarity of expression	0	1	2	Preparedness	0	1	2	Consistency of behaviour	0	1	2	Engagement of pupils	0	1	2	Notes	Interpretation		
Descriptor	Procedural Fidelity	Quality																																																																																																																																																																																																																																																																																																																																										
<b>Activity</b>																																																																																																																																																																																																																																																																																																																																												
Teacher explains the task/activity	Y N	1 2																																																																																																																																																																																																																																																																																																																																										
Teacher checks understanding of the task/activity	Y N	1 2																																																																																																																																																																																																																																																																																																																																										
Teacher reminds pupils that they cannot ask for help	Y N	1 2																																																																																																																																																																																																																																																																																																																																										
<b>Teams</b>																																																																																																																																																																																																																																																																																																																																												
Pupils are in teams of between 3 and 7 (except for special circumstances, e.g. team of 1)	Y N	1 2																																																																																																																																																																																																																																																																																																																																										
Pupils are in clear teams	Y N	1 2																																																																																																																																																																																																																																																																																																																																										
Teams are gender balanced	Y N	1 2																																																																																																																																																																																																																																																																																																																																										
<b>Rules</b>																																																																																																																																																																																																																																																																																																																																												
Rules appropriately verbally reviewed with class	Y N	1 2																																																																																																																																																																																																																																																																																																																																										
Exemplars modelled/described by teacher and/or students	Y N	1 2																																																																																																																																																																																																																																																																																																																																										
Infractions modelled/described by teacher	Y N	1 2																																																																																																																																																																																																																																																																																																																																										
Infractions only described by students	Y N	1 2																																																																																																																																																																																																																																																																																																																																										
Voice level given by teacher	Y N	1 2																																																																																																																																																																																																																																																																																																																																										
<b>Game Specifics</b>																																																																																																																																																																																																																																																																																																																																												
Teacher states when the game begins	Y N																																																																																																																																																																																																																																																																																																																																											
Teacher states how long the game will be played for	Y N																																																																																																																																																																																																																																																																																																																																											
Teacher sets timer	Y N																																																																																																																																																																																																																																																																																																																																											
Teacher states that they will monitor infractions	Y N																																																																																																																																																																																																																																																																																																																																											
Teacher states that 4 infractions are permitted	Y N																																																																																																																																																																																																																																																																																																																																											
Teacher reminds pupils that they are not competing against each other	Y N																																																																																																																																																																																																																																																																																																																																											
Notes	Interpretation																																																																																																																																																																																																																																																																																																																																											
Adaptations	Notes	Interpretation																																																																																																																																																																																																																																																																																																																																										
Teams																																																																																																																																																																																																																																																																																																																																												
Rules																																																																																																																																																																																																																																																																																																																																												
Activity																																																																																																																																																																																																																																																																																																																																												
Game specifics																																																																																																																																																																																																																																																																																																																																												
How do pupils respond to the announcement of the game?	N/A	0 1 2																																																																																																																																																																																																																																																																																																																																										
How attentive are pupils to the teacher's instructions and examples regarding the game?	N/A	0 1 2																																																																																																																																																																																																																																																																																																																																										
How enthusiastic/willing to participate are pupils when discussing the rules?	N/A	0 1 2																																																																																																																																																																																																																																																																																																																																										
Notes	Interpretation																																																																																																																																																																																																																																																																																																																																											
Descriptor	Procedural Fidelity	Quality																																																																																																																																																																																																																																																																																																																																										
<b>Check, Comment, Redirect</b>																																																																																																																																																																																																																																																																																																																																												
Teacher records majority of infractions on scoreboard	Y N	1 2																																																																																																																																																																																																																																																																																																																																										
Teacher identifies majority rule broken	Y N	1 2																																																																																																																																																																																																																																																																																																																																										
Teacher discreetly indicates rules broken to specific pupil most of the time	Y N	1 2																																																																																																																																																																																																																																																																																																																																										
Teacher frequently identifies rule breaking team	Y N	1 2																																																																																																																																																																																																																																																																																																																																										
Rest of team frequently praised for adhering to rules	Y N	1 2																																																																																																																																																																																																																																																																																																																																										
Other teams frequently praised for adhering to rules	Y N	1 2																																																																																																																																																																																																																																																																																																																																										
Teacher does not punish pupils/teams for infractions	Y N																																																																																																																																																																																																																																																																																																																																											
<b>Game Management</b>																																																																																																																																																																																																																																																																																																																																												
Teacher monitors behaviour	Y N	1 2																																																																																																																																																																																																																																																																																																																																										
Teacher does not interact with pupils	Y N																																																																																																																																																																																																																																																																																																																																											
Teacher adheres to time limit	Y N																																																																																																																																																																																																																																																																																																																																											
Teacher announces the end of the game	Y N																																																																																																																																																																																																																																																																																																																																											
Notes	Interpretation																																																																																																																																																																																																																																																																																																																																											
Adaptations	Notes	Interpretation																																																																																																																																																																																																																																																																																																																																										
Check, comment, redirect																																																																																																																																																																																																																																																																																																																																												
Game management																																																																																																																																																																																																																																																																																																																																												
<b>Team</b>	1	2	3	4	5																																																																																																																																																																																																																																																																																																																																							
Rule 1 We will work quietly																																																																																																																																																																																																																																																																																																																																												
Rule 2 We will be polite to others																																																																																																																																																																																																																																																																																																																																												
Rule 3 We will get out of our seats with permission																																																																																																																																																																																																																																																																																																																																												
Rule 4 We will follow directions																																																																																																																																																																																																																																																																																																																																												
TOTAL Infractions																																																																																																																																																																																																																																																																																																																																												
Generally, do rule breaking pupils correct their behaviour following an infraction?	N/A	0 1 2																																																																																																																																																																																																																																																																																																																																										
Generally, how well do pupils respond to a member of their team getting a check?	N/A	0 1 2																																																																																																																																																																																																																																																																																																																																										
Are there any externalising responses from pupils when they receive a check?																																																																																																																																																																																																																																																																																																																																												
Notes	Interpretation																																																																																																																																																																																																																																																																																																																																											
	PIA	P	No																																																																																																																																																																																																																																																																																																																																									
Rules Poster																																																																																																																																																																																																																																																																																																																																												
Voice Levels Poster																																																																																																																																																																																																																																																																																																																																												
Team Assignment Wall-Chart																																																																																																																																																																																																																																																																																																																																												
Scoreboard/Poster																																																																																																																																																																																																																																																																																																																																												
Rules on Desk																																																																																																																																																																																																																																																																																																																																												
GBG Booklets																																																																																																																																																																																																																																																																																																																																												
Timer																																																																																																																																																																																																																																																																																																																																												
Stamps/Stickers for Booklets																																																																																																																																																																																																																																																																																																																																												
Reinforcers																																																																																																																																																																																																																																																																																																																																												
Notes	Interpretation																																																																																																																																																																																																																																																																																																																																											
Descriptor	Procedural Fidelity	Quality																																																																																																																																																																																																																																																																																																																																										
<b>Game Management</b>																																																																																																																																																																																																																																																																																																																																												
Teacher repeats 4 checks or less criterion	Y N	1 2																																																																																																																																																																																																																																																																																																																																										
Teacher announces winning teams only	Y N	1 2																																																																																																																																																																																																																																																																																																																																										
<b>Reinforcers</b>																																																																																																																																																																																																																																																																																																																																												
Members of winning team receive stamp (or marker etc)	Y N	1 2																																																																																																																																																																																																																																																																																																																																										
Star placed on wall-chart	Y N	1 2																																																																																																																																																																																																																																																																																																																																										
Notes	Interpretation																																																																																																																																																																																																																																																																																																																																											
Adaptations	Notes	Interpretation																																																																																																																																																																																																																																																																																																																																										
Game Management																																																																																																																																																																																																																																																																																																																																												
Reinforcers																																																																																																																																																																																																																																																																																																																																												
Generally, what is the level of interest and attentiveness to the reinforcers?	N/A	0 1 2																																																																																																																																																																																																																																																																																																																																										
Generally, how do team leaders respond to sticking items on the board?	N/A	0 1 2																																																																																																																																																																																																																																																																																																																																										
Generally, how do pupils respond if they do not win the game? (note behaviours)	N/A	0 1 2																																																																																																																																																																																																																																																																																																																																										
Notes	Interpretation																																																																																																																																																																																																																																																																																																																																											
<b>Reinforcers</b>																																																																																																																																																																																																																																																																																																																																												
Type	Tangible	Intangible	Token																																																																																																																																																																																																																																																																																																																																									
Praise(s) Given																																																																																																																																																																																																																																																																																																																																												
Delay of Gratification	Immediate	Delayed																																																																																																																																																																																																																																																																																																																																										
Notes	Interpretation																																																																																																																																																																																																																																																																																																																																											
Interest and enthusiasm	0	1	2																																																																																																																																																																																																																																																																																																																																									
Clarity of expression	0	1	2																																																																																																																																																																																																																																																																																																																																									
Preparedness	0	1	2																																																																																																																																																																																																																																																																																																																																									
Consistency of behaviour	0	1	2																																																																																																																																																																																																																																																																																																																																									
Engagement of pupils	0	1	2																																																																																																																																																																																																																																																																																																																																									
Notes	Interpretation																																																																																																																																																																																																																																																																																																																																											

## Appendix 6: Exploratory factor analyses of GBG lesson observation schedule

Both Raykov's composite reliability (Raykov, 1998) and Cronbach alpha were used to assess the internal consistency of the three subscales (fidelity, quality and participant responsiveness). Unlike the widely used Cronbach alpha coefficient, Raykov's composite reliability is not based on  $\tau$ -equivalent, which assumes that the items measure the same construct on the same scale with the same degree of precision (Raykov, 1997). Parallel analysis indicated a two-factor structure for the GBG observation schedule. Subsequently a 2-factor EFA was conducted, results of which are presented in Table A6 below. All items were found to substantially load onto the two domains, although item 6 was found to cross-load onto both factors. Factor 1 concerned procedural fidelity and quality, and factor 2 related to pupil responsiveness.

**Table A6: GBG Observations 2-factor EFA**

	Factor 1 <sup>a</sup>	Factor 2 <sup>b</sup>
1. Pre-game fidelity	0.476***	-0.142
2. During game fidelity	0.475***	0.047
3. Post-game fidelity	0.372***	-0.107
4. How do pupils respond to the announcement of the game?	0.195	0.392**
5. How attentive are pupils to the teacher's instructions and examples regarding the game?	0.354*	0.588***
6. How enthusiastic/willing to participate are pupils when discussing the game?	0.436**	0.568***
7. Do rule breaking pupils correct behaviour following infraction?	-0.010	0.739***
8. How well do pupils respond to members of their team getting a	-0.001	0.840***
9. Teacher interest and enthusiasm	0.732***	0.085
10. Teacher clarity of expression	0.837***	-0.116
11. Teacher preparedness	0.722***	-0.022
12. Teacher consistency of behaviour	0.507***	0.005
13. Teacher engagement of pupils	0.805***	0.107

Note. \*  $p < .05$  \*\*  $p < .01$  \*\*\*  $p < .001$ .

$\chi^2(53) = 116.106$ ,  $p < .001$ ; RMSEA = .109 (.082, .136),  $p < .05$ ; CFI = .880; TLI = .824.

Correlation between factors:  $r = .45$ ,  $p < .05$

a :  $\alpha = .65$  ; Raykov  $\omega = .66$ , b :  $\alpha = .70$  ; Raykov  $\omega = .70$

## Appendix 7: Single sample t-tests comparing trial school characteristics with national averages

Table A7: Primary school sample characteristics and national averages

School characteristic	National average	Total sample (n=77)	Comparison of total sample and national average (one sample t-tests)
<b>Size<sup>1</sup> – number of full-time equivalent (FTE) students on roll</b>	269	306.92 (162.11)	t = 2.05, df = 76, p = 0.44
<b>Attendance<sup>2</sup> – overall absence (% half days)</b>	4.6%	4.21% (0.93)	t = 3.68, df = 76, p < .001
<b>FSM<sup>3</sup> – proportion of students eligible for free school meals</b>	15.6%	25.99% (13.34)	t = 6.83, df = 76, p < .001
<b>EAL<sup>3</sup> – proportion of students speaking English as an additional language</b>	19.4%	22.61% (26.83)	t = 1.05, df = 76, p = .300
<b>SEND<sup>4</sup> – proportion of students with SEND</b>	15.4%	19.50% (7.84)	t = 4.59, df = 76, p < .001
<b>Attainment<sup>5</sup> – proportion of pupil achieving level 4+ in English and maths</b>	80%	75.54 (11.46)	t = 3.42, df = 76, p = .001

<sup>1</sup> DFE (2016). *Schools, Pupils and their Characteristics: January 2016*. London: DFE.

<sup>2</sup> DFE (2016). *Pupil absence in schools in England: 2014 to 2015*. London: DFE.

<sup>3</sup> DFE (2015). *Schools, Pupils and their Characteristics*. London: DFE.

<sup>4</sup> DFE (2015). *Special educational needs in England: January 2015*. London: DFE.

<sup>5</sup> DFE (2015). *National curriculum assessments at key stage 2 in England, 2015 (revised)*. London: DFE.

## Appendix 8: Predictors of missing data at follow-up

Table A8: Logistic regression to predict missingness

$\beta_{0ij} = 0.562 (0.0889)$		
	Co-efficient $\beta$ (SE)	p
<b>School level</b>	0.002 (0.001)	.049
Trial group (if GBG)	-0.005 (0.016)	.757
<b>Pupil level</b>	0.125 (0.003)	<.001
Gender (if male)	-0.006 (0.014)	.803
FSM (if eligible)	0.020 (0.016)	.211
KS1 READ_POINTS	-0.016 (0.002)	<.001
Conduct problems	0.015 (0.008)	.060
Concentration problems	-0.019 (0.009)	.035
Disruptive behaviour	0.000 (0.018)	.999
Pro-social	-0.026 (0.012)	.030
-2*Loglikelihood =	2254.310	
	$X^2$ (df=8, n=2924) = 622.240, p <.001	



## Appendix 9: MLM ITT and sub-group analyses

Table A9: Reading – complete case

	Empty model		Model 1.1		Model 1.2		Model 1.3		Model 1.4		Model 1.5	
	$\beta_{0ij} = 32.883 (0.402)$		$\beta_{0ij} = -0.759 (0.733)$		$\beta_{0ij} = 3.129 (1.004)$		$\beta_{0ij} = -0.659 (0.793)$		$\beta_{0ij} = -0.347 (0.746)$		$\beta_{0ij} = 3.377 (1.032)$	
	Co-efficient $\beta$	p	Co-efficient $\beta$	p	Co-efficient $\beta$	p	Co-efficient $\beta$	p	Co-efficient $\beta$	p	Co-efficient $\beta$	p
	(SE)		(SE)		(SE)		(SE)		(SE)		(SE)	
<b>School</b>	9.070 (1.994)	<.001	5.987 (1.197)	<.001	3.318 (0.760)	<.001	5.994 (1.199)	<.001	5.746 (1.158)	<.001	3.302 (0.759)	<.001
FSM					-0.119 (0.019)	<.001					-0.121 (0.019)	<.001
School size					-0.001 (0.001)	.500					-0.001 (0.001)	.500
<b>Trial group (if GBG)</b>			<b>0.331 (0.622)</b>	<b>.299</b>	<b>0.583 (0.500)</b>	<b>.143</b>	<b>0.796 (0.690)</b>	<b>.127</b>	<b>0.442 (0.656)</b>	<b>.252</b>	<b>0.638 (0.616)</b>	<b>.151</b>
<b>Pupil</b>	97.936 (2.802)	<.001	41.581 (1.194)	<.001	41.536 (1.192)	<.001	41.205 (1.192)	<.001	41.516 (1.192)	<.001	41.046 (1.187)	<.001
Gender (if male)					-0.062 (0.264)	.409	0.192 (0.400)	.317			-0.077 (0.439)	.429
FSM (if eligible)					-0.487 (0.321)	.064			-0.961 (0.469)	.020	-1.581 (0.696)	.012
Risk status (if at risk)							-2.342 (1.112)	.017			-2.644 (1.136)	<.001
T1 score			2.065 (0.036)	<.001	2.045 (0.036)	<.001	2.059 (0.037)	<.001	2.052 (0.036)	<.001	2.040 (0.037)	<.001
<b>Trial group*FSM</b>									<b>0.501 (0.637)</b>	<b>.215</b>	<b>1.753 (0.915)</b>	<b>.027</b>
Trial group*Gender							-0.878 (0.569)	.062			-0.357 (0.638)	.289
Trial group*Risk status							0.191 (1.396)	.444			-0.100 (1.398)	.472
FSM*Gender											1.064 (0.929)	.125
FSM*Risk status											1.496 (1.028)	.072
Gender*risk status							1.868 (1.282)	.072			1.735 (1.278)	.087
Trial group*FSM*Gender											-2.138 (1.331)	.054
<b>Trial group*Gender*Risk status</b>							<b>0.451 (1.658)</b>	<b>.394</b>			<b>1.212 (1.761)</b>	<b>.245</b>
Trial group*FSM*Gender*Risk status											-1.295 (1.711)	.218
-2*Loglikelihood =	18800.200		16569.530		16532.620		16297.478		16563.155		16253.656	
			X <sup>2</sup> (df=2, n=2504) =		X <sup>2</sup> (df=6, n=2504) =		X <sup>2</sup> (df=8, n=2466) =		X <sup>2</sup> (df=4 n=2504) =		X <sup>2</sup> (df=15, n=2466) =	
			2230.670, p <.001		2267.580, p <.001		2502.722, p <.001		2237.045, p <.001		2546.544, p <.001	

Table A10: Reading – MI

	Model 2.1		Model 2.2		Model 2.3		Model 2.4		Model 2.5	
	$\beta_{0ij} = -0.064 (0.741)$		$\beta_{0ij} = 4.005 (0.973)$		$\beta_{0ij} = -0.243 (0.841)$		$\beta_{0ij} = 0.0583 (0.712)$		$\beta_{0ij} = 4.303 (1.029)$	
	Co-efficient $\beta$	p	Co-efficient $\beta$	p	Co-efficient $\beta$	p	Co-efficient $\beta$	p	Co-efficient $\beta$	p
	(SE)		(SE)		(SE)		(SE)		(SE)	
<b>School</b>	6.093 (1.264)	<.001	3.488 (0.828)	<.001	6.031 (1.201)	<.001	5.607 (1.146)	<.001	3.235 (0.746)	<.001
FSM			-0.119 (0.020)	<.001					-0.121 (0.019)	<.001
School size			-0.001 (0.002)	.309					-0.001 (0.001)	.500
<b>Trial group (if GBG)</b>	<b>0.344 (0.637)</b>	<b>.295</b>	<b>0.576 (0.512)</b>	<b>.132</b>	<b>0.715 (0.692)</b>	<b>.153</b>	<b>0.456 (0.664)</b>	<b>.246</b>	<b>0.642 (0.611)</b>	<b>.149</b>
<b>Pupil</b>	43.757 (1.399)	<.001	43.313 (1.179)	<.001	42.547 (1.297)	<.001	43.470 (1.288)	<.001	42.397 (1.272)	<.001
Gender (if male)			-0.168 (0.273)	.268	0.153 (0.411)	.356			-0.121 (0.433)	.390
FSM (if eligible)			-0.596 (0.336)	.038			-1.009 (0.456)	.014	-1.394 (0.706)	.024
Risk status (if at risk)					-1.785 (1.040)	.043			-2.317 (1.247)	<.001
T1 score	2.021 (0.037)	<.001	1.994 (0.034)	<.001	2.034 (0.039)	<.001	1.995 (0.034)	<.001	1.990 (0.037)	<.001
<b>Trial group*FSM</b>							<b>0.439 (0.680)</b>	<b>.258</b>	<b>1.317 (0.883)</b>	<b>.068</b>
Trial group*Gender					-0.649 (0.580)	.131			-0.311 (0.613)	.305
Trial group*Risk status					-0.091 (1.323)	.468			0.023 (1.493)	.492
FSM*Gender									0.826 (0.948)	.192
FSM*Risk status									1.088 (0.989)	.136
Gender*risk status					1.116 (1.169)	.171			1.041 (1.331)	.218
Trial group*									-1.653 (1.339)	.109
FSM*Gender										
<b>Trial group*</b>					<b>0.607 (1.521)</b>	<b>.345</b>			<b>1.042 (1.708)</b>	<b>.271</b>
<b>Gender*Risk status</b>										
Trial group*FSM*									-0.637 (1.724)	.356
Gender*Risk status										

Table A11: Concentration problems – complete case

	Empty model		Model 1.1		Model 1.2		Model 1.3		Model 1.4		Model 1.5	
	$\beta_{0ij} = 2.520 (0.046)$		$\beta_{0ij} = 0.807 (0.080)$		$\beta_{0ij} = 0.488 (0.145)$		$\beta_{0ij} = 0.802 (0.082)$		$\beta_{0ij} = 0.791 (0.079)$		$\beta_{0ij} = 0.538 (0.147)$	
	Co-efficient $\beta$	p	Co-efficient $\beta$	p	Co-efficient $\beta$	p	Co-efficient $\beta$	p	Co-efficient $\beta$	p	Co-efficient $\beta$	p
	(SE)		(SE)		(SE)		(SE)		(SE)		(SE)	
<b>School</b>	0.120 (0.026)	<.001	0.165 (0.031)	<.001	0.144 (0.027)	<.001	0.160 (0.030)	<.001	0.158 (0.029)	<.001	0.143 (0.027)	<.001
FSM					0.004 (0.004)	.160					0.004 (0.004)	.160
School size					0.000 (0.000)	.500					0.000 (0.000)	.500
<b>Trial group (if GBG)</b>			<b>0.035 (0.100)</b>	<b>.364</b>	<b>0.033 (0.095)</b>	<b>.364</b>	<b>0.057 (0.105)</b>	<b>.295</b>	<b>0.029 (0.100)</b>	<b>.386</b>	<b>0.046 (0.105)</b>	<b>.331</b>
<b>Pupil</b>	1.165 (0.033)	<.001	0.659 (0.019)	<.001	0.639 (0.019)	<.001	0.634 (0.018)	<.001	0.654 (0.019)	<.001	0.629 (0.018)	<.001
Gender (if male)					0.262 (0.034)	<.001	0.246 (0.049)	<.001			0.252 (0.054)	<.001
FSM (if eligible)					0.185 (0.040)	<.001			0.192 (0.058)	<.001	0.202 (0.086)	.009
Risk status (if at risk)							0.324 (0.137)	.009			0.318 (0.142)	.013
T1 score			0.659 (0.015)	<.001	0.617 (0.016)	<.001	0.587 (0.017)	<.001	0.649 (0.016)	<.001	0.577 (0.017)	<.001
<b>Trial group*FSM</b>									<b>-0.021 (0.080)</b>	<b>.397</b>	<b>-0.043 (0.114)</b>	<b>.352</b>
Trial group*Gender							0.022 (0.071)	.378			0.012 (0.080)	.440
Trial group*Risk status							0.045 (0.174)	.397			0.042 (0.175)	.405
FSM*Gender											-0.048 (0.111)	.334
FSM*Risk status											-0.024 (0.124)	.425
Gender*risk status							0.082 (0.153)	.295			0.100 (0.153)	.258
Trial group*											0.110 (0.163)	.251
FSM*Gender												
<b>Trial group*</b>							<b>-0.313 (0.205)</b>	<b>.063</b>			<b>-0.317 (0.218)</b>	<b>.074</b>
<b>Gender*Risk status</b>												
Trial group*FSM*											-0.042 (0.210)	.421
Gender*Risk status												
-2*Loglikelihood =	7619.843		6138.592		6042.102		6041.564		6104.210		6002.396	
			X <sup>2</sup> (df=2, n=2469) =		X <sup>2</sup> (df=6, n=2463) =		X <sup>2</sup> (df=8, n=2468) =		X <sup>2</sup> (df=4 n=2463) =		X <sup>2</sup> (df=15, n=2462) =	
			1481.251, p <.001		1577.741, p <.001		1578.279, p <.001		1515.633, p <.001		1617.447, p <.001	

Table A12: Concentration problems – MI

	Model 2.1		Model 2.2		Model 2.3		Model 2.4		Model 2.5	
	$\beta_{0ij} = 0.834 (0.080)$		$\beta_{0ij} = 0.503 (0.146)$		$\beta_{0ij} = 0.842 (0.086)$		$\beta_{0ij} = 0.810 (0.079)$		$\beta_{0ij} = 0.547 (0.148)$	
	Co-efficient $\beta$ (SE)	p	Co-efficient $\beta$ (SE)	p	Co-efficient $\beta$ (SE)	p	Co-efficient $\beta$ (SE)	p	Co-efficient $\beta$ (SE)	p
<b>School</b>	0.162 (0.029)	<.001	0.148 (0.028)	<.001	0.156 (0.029)	<.001	0.159 (0.030)	<.001	0.141 (0.026)	<.001
FSM			0.005 (0.004)	.108					0.005 (0.004)	.108
School size			0.000 (0.000)	.500					0.000 (0.000)	.500
<b>Trial group (if GBG)</b>	<b>0.022 (0.099)</b>	<b>.413</b>	<b>0.025 (0.095)</b>	<b>.398</b>	<b>0.045 (0.104)</b>	<b>.334</b>	<b>0.018 (0.100)</b>	<b>.429</b>	<b>0.052 (0.103)</b>	<b>.309</b>
<b>Pupil</b>	0.670 (0.021)	<.001	0.655 (0.018)	<.001	0.644 (0.018)	<.001	0.671 (0.020)	<.001	0.642 (0.019)	<.001
Gender (if male)			0.271 (0.034)	<.001	0.256 (0.048)	<.001			0.270 (0.053)	<.001
FSM (if eligible)			0.192 (0.041)	<.001			0.196 (0.053)	<.001	0.206 (0.081)	.006
Risk status (if at risk)					0.363 (0.136)	<.001			0.348 (0.134)	.005
T1 score	0.653 (0.015)	<.001	0.611 (0.015)	<.001	0.574 (0.021)	<.001	0.643 (0.015)	<.001	0.569 (0.018)	<.001
<b>Trial group*FSM</b>							<b>-0.007 (0.083)</b>	<b>.469</b>	<b>-0.044 (0.112)</b>	<b>.348</b>
Trial group*Gender					0.023 (0.068)	.367			-0.004 (0.080)	.480
Trial group*Risk status					-0.043 (0.178)	.405			0.002 (0.170)	.496
FSM*Gender									-0.036 (0.106)	.367
FSM*Risk status									-0.037 (0.121)	.378
Gender*risk status					0.044 (0.158)	.390			0.071 (0.144)	.312
Trial group*									0.091 (0.158)	.281
FSM*Gender										
<b>Trial group*</b>					<b>-0.197 (0.206)</b>	<b>.169</b>			<b>-0.248 (0.198)</b>	<b>.106</b>
<b>Gender*Risk status</b>										
Trial group*FSM*									0.002 (0.204)	.496
Gender*Risk status										

Table A13: Disruptive behaviour - complete case

	<b>Empty model</b>		<b>Model 1.1</b>		<b>Model 1.2</b>		<b>Model 1.3</b>		<b>Model 1.4</b>		<b>Model 1.5</b>	
	$\beta_{0ij} = 1.710 (0.033)$		$\beta_{0ij} = 0.501 (0.053)$		$\beta_{0ij} = 0.400 (0.100)$		$\beta_{0ij} = 0.579 (0.059)$		$\beta_{0ij} = 0.498 (0.053)$		$\beta_{0ij} = 0.529 (0.104)$	
	<b>Co-efficient <math>\beta</math></b>	<b>p</b>	<b>Co-efficient <math>\beta</math></b>	<b>p</b>	<b>Co-efficient <math>\beta</math></b>	<b>p</b>	<b>Co-efficient <math>\beta</math></b>	<b>p</b>	<b>Co-efficient <math>\beta</math></b>	<b>p</b>	<b>Co-efficient <math>\beta</math></b>	<b>p</b>
	<b>(SE)</b>		<b>(SE)</b>		<b>(SE)</b>		<b>(SE)</b>		<b>(SE)</b>		<b>(SE)</b>	
<b>School</b>	0.060 (0.013)	<.001	0.066 (0.013)	<.001	0.066 (0.013)	<.001	0.066 (0.013)	<.001	0.066 (0.013)	<.001	0.064 (0.013)	<.001
FSM					0.001 (0.002)	.309					0.001 (0.002)	.309
School size					0.000 (0.000)	.500					0.000 (0.000)	.500
<b>Trial group (if GBG)</b>			<b>0.051 (0.065)</b>	<b>.219</b>	<b>0.055 (0.065)</b>	<b>.199</b>	<b>0.008 (0.071)</b>	<b>.456</b>	<b>0.032 (0.066)</b>	<b>.316</b>	<b>-0.020 (0.073)</b>	<b>.394</b>
<b>Pupil</b>	0.659 (0.019)	<.001	0.371 (0.011)	<.001	0.365 (0.011)	<.001	0.364 (0.011)	<.001	0.369 (0.011)	<.001	0.361 (0.101)	<.001
Gender (if male)					0.132 (0.026)	<.001	0.065 (0.037)	.039			0.057 (0.041)	.082
FSM (if eligible)					0.084 (0.030)	.003			0.045 (0.043)	.147	0.042 (0.065)	.258
Risk status (if at risk)							0.122 (0.109)	.131			0.136 (0.112)	.113
T1 score			0.717 (0.017)	<.001	0.689 (0.017)	<.001	0.623 (0.026)	<.001	0.712 (0.017)	<.001	0.620 (0.026)	<.001
<b>Trial group*FSM</b>									<b>0.069 (0.060)</b>	<b>.125</b>	<b>0.088 (0.086)</b>	<b>.154</b>
Trial group*Gender							0.132 (0.054)	.007			0.159 (0.060)	.004
Trial group*Risk status							0.051 (0.132)	.356			0.038 (0.132)	.386
FSM*Gender											0.015 (0.085)	.429
FSM*Risk status											-0.042 (0.094)	.326
Gender*risk status							0.188 (0.116)	.053			0.194 (0.116)	.048
Trial group*											-0.070 (0.124)	.288
FSM*Gender												
<b>Trial group*</b>							<b>-0.245 (0.155)</b>	<b>.053</b>			<b>-0.322 (0.165)</b>	<b>.026</b>
<b>Gender*Risk status</b>												
Trial group*FSM*											0.219 (0.159)	.084
Gender*Risk status												
-2*Loglikelihood =	6182.421		4698.006		4647.809		4647.789		4673.062		4618.460	

## Good Behaviour Game

X <sup>2</sup> (df=2, n=2469) = 1484.415, p <.001	X <sup>2</sup> (df=6, n=2463) = 1534.612, p <.001	X <sup>2</sup> (df=8, n=2468) = 1534.632, p <.001	X <sup>2</sup> (df=4, n=2463) = 1509.359, p <.001	X <sup>2</sup> (df=15, n=2462) = 1563.961, p <.001
--	--	--	--	---

**Table A14: Disruptive behaviour – MI**

	Model 2.1		Model 2.2		Model 2.3		Model 2.4		Model 2.5	
	$\beta_{0ij} = 0.531 (0.054)$		$\beta_{0ij} = 0.428 (0.106)$		$\beta_{0ij} = 0.608 (0.062)$		$\beta_{0ij} = 0.527 (0.055)$		$\beta_{0ij} = 0.545 (0.106)$	
	Co-efficient $\beta$ (SE)	p	Co-efficient $\beta$ (SE)	p	Co-efficient $\beta$ (SE)	p	Co-efficient $\beta$ (SE)	p	Co-efficient $\beta$ (SE)	p
<b>School</b>	0.074 (0.016)	<.001	0.073 (0.015)	<.001	0.071 (0.015)	<.001	0.072 (0.014)	<.001	0.068 (0.013)	<.001
FSM			0.001 (0.003)	.371					0.001 (0.003)	.371
School size			0.000 (0.000)	.500					0.000 (0.000)	.500
<b>Trial group (if GBG)</b>	<b>0.047 (0.067)</b>	<b>.243</b>	<b>0.048 (0.069)</b>	<b>.243</b>	<b>0.028 (0.073)</b>	<b>.360</b>	<b>0.035 (0.068)</b>	<b>.306</b>	<b>-0.006 (0.073)</b>	<b>.468</b>
<b>Pupil</b>	0.388 (0.011)	<.001	0.378 (0.013)	<.001	0.376 (0.011)	<.001	0.379 (0.011)	<.001	0.372 (0.011)	<.001
Gender (if male)			0.138 (0.025)	<.001	0.087 (0.040)	.015			0.075 (0.041)	.034
FSM (if eligible)			0.091 (0.029)	<.001			0.051 (0.044)	.123	0.042 (0.064)	.255
Risk status (if at risk)					0.178 (0.134)	.092			0.169 (0.108)	.059
T1 score	0.698 (0.016)	<.001	0.669 (0.017)	<.001	0.594 (0.029)	<.001	0.694 (0.017)	<.001	0.597 (0.024)	<.001
<b>Trial group*FSM</b>							<b>0.057 (0.061)</b>	<b>.176</b>	<b>0.069 (0.084)</b>	<b>.206</b>
Trial group*Gender					0.099 (0.055)	.036			0.122 (0.056)	.015
Trial group*Risk status					0.028 (0.159)	.429			0.061 (0.122)	.309
FSM*Gender									0.014 (0.084)	.433
FSM*Risk status									-0.013 (0.091)	.444
Gender*risk status					0.163 (0.120)	.087			0.163 (0.110)	.070
Trial group*									-0.051 (0.118)	.334
FSM*Gender										
<b>Trial group*</b>					<b>-0.240 (0.163)</b>	<b>.071</b>			<b>-0.275 (0.159)</b>	<b>.042</b>
<b>Gender*Risk status</b>										
Trial group*FSM*									0.131 (0.147)	.187
Gender*Risk status										

Good Behaviour Game

Table A15: Pro-social behaviour - complete cases

	Empty model		Model 1.1		Model 1.2		Model 1.3		Model 1.4		Model 1.5	
	$\beta_{0ij} = 4.877 (0.051)$		$\beta_{0ij} = 2.358 (0.117)$		$\beta_{0ij} = 2.795 (0.178)$		$\beta_{0ij} = 3.035 (0.135)$		$\beta_{0ij} = 2.453 (0.119)$		$\beta_{0ij} = 3.277 (0.147)$	
	Co-efficient $\beta$	p	Co-efficient $\beta$	p	Co-efficient $\beta$	p	Co-efficient $\beta$	p	Co-efficient $\beta$	p	Co-efficient $\beta$	p
	(SE)		(SE)		(SE)		(SE)		(SE)		(SE)	
<b>School</b>	0.170 (0.032)	<.001	0.177 (0.032)	<.001	0.170 (0.031)	<.001	0.173 (0.031)	<.001	0.174 (0.032)	<.001	0.167 (0.030)	<.001
FSM					-0.003 (0.004)	.228					-0.003 (0.004)	.228
School size					-0.000 (0.000)	.500					-0.000 (0.000)	.500
<b>Trial group (if GBG)</b>			<b>-0.113 (0.102)</b>	<b>.135</b>	<b>-0.113 (0.101)</b>	<b>.135</b>	<b>-0.085 (0.107)</b>	<b>.216</b>	<b>-0.127 (0.103)</b>	<b>.111</b>	<b>-0.094 (0.108)</b>	<b>.192</b>
<b>Pupil</b>	0.729 (0.021)	<.001	0.551 (0.016)	<.001	0.540 (0.016)	<.001	0.527 (0.015)	<.001	0.547 (0.016)	<.001	0.523 (0.015)	<.001
Gender (if male)					-0.178 (0.031)	<.001	-0.109 (0.044)	.008			-0.109 (0.049)	.013
FSM (if eligible)					-0.143 (0.037)	<.001					-0.184 (0.078)	.009
Risk status (if at risk)							-0.541 (0.126)	<.001	-0.184 (0.053)	<.001	-0.525 (0.131)	<.001
T1 score			0.521 (0.019)	<.001	0.491 (0.019)	<.001	0.412 (0.022)	<.001	0.511 (0.019)	<.001	0.402 (0.022)	<.001
<b>Trial group*FSM</b>									<b>0.082 (0.073)</b>	<b>.131</b>	<b>0.098 (0.104)</b>	<b>.173</b>
Trial group*Gender							-0.111 (0.065)	.044			-0.109 (0.104)	.147
Trial group*Risk status							0.143 (0.159)	.184			0.139 (0.160)	.192
FSM*Gender											0.037 (0.102)	.359
FSM*Risk status											-0.012 (0.113)	.456
Gender*risk status							0.002 (0.140)	.496			0.006 (0.140)	.484
Trial group*											-0.066 (0.149)	.330
FSM*Gender												
<b>Trial group*</b>							<b>0.100 (0.187)</b>	<b>.298</b>			<b>0.124 (0.199)</b>	<b>.268</b>
<b>Gender*Risk status</b>												
Trial group*FSM*											-0.024 (0.191)	.450
Gender*Risk status												
-2*Loglikelihood =	6493.474		5713.173		5649.197		5601.743		5682.954		5570.216	
			X <sup>2</sup> (df=2, n=2469) =		X <sup>2</sup> (df=6, n=2463) =		X <sup>2</sup> (df=8, n=2468) =		X <sup>2</sup> (df=4, n=2463) =		X <sup>2</sup> (df=15, n=2462) =	
			780.301, p <.001		844.277, p <.001		=891.731, p <.001		810.520, p <.001		=923.258, p <.001	

Table A16: Pro-social behaviour – MI

	Model 2.1		Model 2.2		Model 2.3		Model 2.4		Model 2.5	
	$\beta_{0ij} = 2.415 (0.115)$		$\beta_{0ij} = 2.849 (0.182)$		$\beta_{0ij} = 3.058 (0.127)$		$\beta_{0ij} = 2.484 (0.119)$		$\beta_{0ij} = 3.322 (0.187)$	
	Co-efficient $\beta$ (SE)	p	Co-efficient $\beta$ (SE)	p	Co-efficient $\beta$ (SE)	p	Co-efficient $\beta$ (SE)	p	Co-efficient $\beta$ (SE)	p
<b>School</b>	0.182 (0.033)	<.001	0.172 (0.031)	<.001	0.175 (0.031)	<.001	0.175 (0.031)	<.001	0.173 (0.031)	<.001
FSM			-0.003 (0.004)	.228					-0.003 (0.004)	.228
School size			0.000 (0.000)	.500					0.000 (0.000)	.500
<b>Trial group (if GBG)</b>	<b>-0.118 (0.102)</b>	<b>.125</b>	<b>-0.105 (0.101)</b>	<b>.151</b>	<b>-0.071 (0.107)</b>	<b>.256</b>	<b>-0.122 (0.102)</b>	<b>.117</b>	<b>-0.177 (0.110)</b>	<b>.056</b>
<b>Pupil</b>	0.555 (0.016)	<.001	0.549 (0.015)	<.001	0.536 (0.015)	<.001	0.558 (0.016)	<.001	0.530 (0.016)	<.001A
Gender (if male)			-0.177 (0.032)	<.001	-0.112 (0.044)	.005			-0.130 (0.047)	.003
FSM (if eligible)			-0.136 (0.036)	<.001			-0.180 (0.053)	<.001	-0.188 (0.083)	.012
Risk status (if at risk)					-0.503 (0.119)	<.001			-0.518 (0.124)	<.001
T1 score	0.509 (0.018)	<.001	0.479 (0.018)	<.001	0.405 (0.020)	<.001	0.503 (0.019)	<.001	0.397 (0.021)	<.001
<b>Trial group*FSM</b>							<b>0.074 (0.071)</b>	<b>.149</b>	<b>0.104 (0.103)</b>	<b>.156</b>
Trial group*Gender					-0.083 (0.063)	.093			-0.063 (0.070)	.184
Trial group*Risk status					0.081 (0.146)	.291			0.103 (0.145)	.239
FSM*Gender									0.041 (0.100)	.341
FSM*Risk status									-0.012 (0.119)	.460
Gender*risk status					0.012 (0.133)	.464			0.016 (0.130)	.452
Trial group*									-0.074 (0.149)	.309
FSM*Gender										
<b>Trial group*</b>					<b>0.123 (0.179)</b>	<b>.245</b>			<b>0.100 (0.188)</b>	<b>.298</b>
<b>Gender*Risk status</b>										
Trial group*FSM*									0.031 (0.193)	.436
Gender*Risk status										



## Appendix 10: Analyses of teacher-level outcomes

**Table A17: Regression analysis for teacher outcomes with FIML**

Outcome	B	SE	$\beta$	<i>p</i>	Satorra-Bentler chi-square (df) GBG vs UP
<b>Self-efficacy Follow-up</b>					
Self-efficacy at baseline	.409	.061	.483	<.001	1.266 (1)
Trial arm	-.127	.107	-.069	.236	
<b>Stress Follow-up</b>					
Stress at baseline	.557	.067	.531	<.001	1.531 (1)
Trial arm	.076	.089	.049	.394	
<b>Retention Follow-up</b>					
Retention at baseline	.546	.098	.509	<.001	3.383 (1)
Trial arm	.063	.154	.023	.683	

Data on the variables of interest were missing (15.1-16.8%) completely at random, allowing for the use of all available information ( $n=279$ ). To account for missingness, subsequent models were tested through Full information Maximum Likelihood (FIML) with robust standard errors (MLR). Given that all models were saturated (i.e. number of estimated parameters equals the number of data points), model fit was not evaluated. Multigroup analyses were conducted for each model to compare each pathway between GBG and UP groups. An unconstrained model in which each path was allowed to be unequal between groups was compared to a nested model in which the path was constrained to equality. Difference testing was explored through Satorra-Bentler chi-square, where a statistically significant chi-square indicated non-invariance (the pathway differs between the groups). Results for each model showed that the effect of trial group was non-significant.

## Appendix 11: Implementation analyses

Table A18: Reading 2015/16

	Empty model		Model 1.1		Model 1.2		Model 2.1		Model 2.2	
	$\beta_{0ij} = 32.195 (0.567)$		$\beta_{0ij} = 2.987 (1.129)$		$\beta_{0ij} = 2.720 (1.153)$		$\beta_{0ij} = 0.567 (5.972)$		$\beta_{0ij} = -0.364 (5.865)$	
	Co-efficient $\beta$	p	Co-efficient $\beta$ (SE)	p	Co-efficient $\beta$ (SE)	p	Co-efficient $\beta$ (SE)	p	Co-efficient $\beta$ (SE)	p
<b>Class</b>	13.120 (3.370)	<.001	3.503 (1.129)	<.001	4.233 (1.181)	<.001	6.305 (1.724)	<.001	7.174 (1.728)	<.001
Procedural Fidelity and Quality (compared to low)			-2.820 (1.137) – if mod	.009	-3.634 (1.068) – if mod	<.001	0.019 (0.042)	.328	-0.013 (0.041)	.376
Participant responsiveness (compared to low)			-1.179 (1.324) - if high	.190	-1.935 (1.317) – if high	.076				
Participant reach (compared to low)			-0.980 (0.985) – if mod	.165	-0.674 (0.971) - if mod	.248	0.019 (0.025)	.227	0.013 (0.025)	.303
Dosage (compared to low)			1.936 (1.275) – if high	.069	2.129 (1.371) – if high	.066				
			-0.587 (1.191) – if mod	.314	0.478 (1.153) - if mod	.342	-0.032 (0.057)	.290	0.017 (0.053)	.376
			-0.148 (0.925) – if high	.437	0.976 (0.879) – if high	.140				
			-0.216 (0.988) – if mod	.414	0.560 (1.019) – if mod	.293	0.000 (0.001)	.500	0.000 (0.001)	.500
			2.557 (1.245) – if high	.025	3.034 (1.238) – if high	.010				
<b>Pupil</b>	90.764 (3.875)	<.001	37.501 (1.762)	<.001	38.863 (1.640)	<.001	37.494 (1.762)	<.001	39.355 (1.656)	<.001
Gender (if male)			-0.203 (0.408)	.309	-0.374 (0.368)	.154	-0.206 (0.408)		-0.215 (0.360)	.274
FSM (if eligible)			-0.713 (0.474)	.067	-0.658 (0.440)		-0.754 (0.475)		-0.727 (0.455)	.055
T1 score			2.017 (0.057)	<.001	1.975 (0.056)	<.001	2.024 (0.057)	<.001	1.968 (0.054)	<.001
-2*Loglikelihood =	8538.311		6187.482				6206.159			
			X <sup>2</sup> (df=15, n=950) =2350.829, p <.001				X <sup>2</sup> (df=7, n=950) = 7455.521, p <.001			

**Table A19: Concentration problems 2015/16**

	Empty model		Model 1.1		Model 1.2		Model 2.1		Model 2.2	
	$\beta_{0ij} = 2.531 (0.059)$		$\beta_{0ij} = 0.686 (0.291)$		$\beta_{0ij} = 0.856 (0.274)$		$\beta_{0ij} = 0.814 (0.938)$		$\beta_{0ij} = 0.894 (0.869)$	
	Co-efficient $\beta$	p	Co-efficient $\beta$ (SE)	p	Co-efficient $\beta$ (SE)	p	Co-efficient $\beta$ (SE)	p	Co-efficient $\beta$ (SE)	p
<b>Class</b>	0.128 (0.036)	<.001	0.161 (0.041)	<.001	0.187 (0.043)	<.001	0.174 (0.044)	<.001	0.191 (0.043)	<.001
Procedural Fidelity and Quality (compared to low)			0.156 (0.217) – if mod	.238	0.145 (0.201) – if mod	.238	-0.002 (0.007)	.387	0.003 (0.006)	.310
			-0.012 (0.251) - if high	.480	-0.004 (0.246) – if high	.492				
Participant responsiveness (compared to low)			-0.049 (0.184) – if mod	.394	-0.185 (0.177) - if mod	.150	-0.001 (0.004)	.402	-0.003 (0.004)	.229
			-0.240 (0.242) – if high	.162	-0.266 (0.241) – if high	.139				
Participant reach (compared to low)			0.036 (0.231) – if mod	.437	0.074 (0.229) - if mod	.375	0.000 (0.009)	.500	-0.002 (0.008)	.402
			-0.028 (0.175) – if high	.437	-0.094 (0.167) – if high	.289				
Dosage (compared to low)			0.021 (0.186) – if mod	.457	0.046 (0.177) – if mod	.398	0.000 (0.000)	.500	0.000 (0.000)	.500
			-0.069 (0.238) – if high	.387	0.022 (0.277) – if high	.468				
<b>Pupil</b>	1.127 (0.049)	<.001	0.633 (0.030)	<.001	0.630 (0.028)	<.001	0.632 (0.030)	<.001	0.631 (0.027)	<.001
Gender (if male)			0.244 (0.055)	<.001	0.232 (0.048)	<.001	0.245 (0.055)	<.001	0.235 (0.050)	<.001
FSM (if eligible)			0.204 (0.063)	<.001	0.171 (0.054)	<.001	0.210 (0.063)	<.001	0.194 (0.056)	<.001
T1 score			0.637 (0.026)	<.001	0.615 (0.024)	<.001	0.636 (0.026)	<.001	0.614 (0.024)	<.001
-2*Loglikelihood =	3364.276		2279.005				2281.486			
			X <sup>2</sup> (df=15, n=924) =1085.271, p <.001				X <sup>2</sup> (df=7, n=924) = 1082.790, p <.001			

Table A20: Disruptive behaviour 2015/16

	Empty model		Model 1.1		Model 1.2		Model 2.1		Model 2.2	
	$\beta_{0ij} = 1.741 (0.042)$		$\beta_{0ij} = 0.156 (0.197)$		$\beta_{0ij} = 0.319 (0.191)$		$\beta_{0ij} = -0.878 (0.664)$		$\beta_{0ij} = -0.364 (0.607)$	
	Co-efficient $\beta$	p	Co-efficient $\beta$ (SE)	p	Co-efficient $\beta$ (SE)	p	Co-efficient $\beta$ (SE)	p	Co-efficient $\beta$ (SE)	p
<b>Class</b>	0.061 (0.018)	<.001	0.067 (0.018)	<.001	0.074 (0.018)	<.001	0.083 (0.022)	<.001	0.083 (0.022)	<.001
Procedural Fidelity and Quality (compared to low)			-0.175 (0.145) – if mod	.118	-0.067 (0.135) – if mod	.310	-0.002 (0.005)	.346	-0.001 (0.004)	.402
Participant responsiveness (compared to low)			-0.039 (0.168) - if high	.410	0.045 (0.166) – if high	.394				
Participant reach (compared to low)			-0.085 (0.124) – if mod	.248	-0.125 (0.121) – if mod	.155	-0.002 (0.003)	.253	-0.001 (0.003)	.372
Dosage (compared to low)			-0.131 (0.163) – if high	.215	-0.206 (0.163) – if high	.107				
			0.571 (0.155) – if mod	<.001	0.450 (0.147) – if mod	<.001	0.014 (0.006)	.012	0.010 (0.006)	.051
			0.316 (0.117) – if high	.005	0.232 (0.108) – if mod	.019				
			0.174 (0.124) – if mod	.086	0.078 (0.112) – if mod	.244	0.000 (0.000)	.500	0.000 (0.000)	.500
			0.175 (0.160) – if high	.142	0.077 (0.142) – if high	.296				
<b>Pupil</b>	0.675 (0.029)	<.001	0.390 (0.019)	<.001	0.386 (0.017)	<.001	0.390 (0.019)	<.001	0.389 (0.017)	<.001
Gender (if male)			0.179 (0.043)	<.001	0.202 (0.040)	<.001	0.180 (0.043)	<.001	0.189 (0.039)	<.001
FSM (if eligible)			0.100 (0.049)	.023	0.108 (0.043)	.006	0.109 (0.049)	.013	0.101 (0.043)	.009
T1 score			0.735 (0.030)	<.001	0.699 (0.032)	<.001	0.734 (0.030)	<.001	0.697 (0.027)	<.001
-2*Loglikelihood =	2783.248		1818.569				1824.912			
			X <sup>2</sup> (df=15, n=924) =946.679, p <.001				X <sup>2</sup> (df=7, n=924) = 958.336, p <.001			

Table A21: Pro-social behaviour 2015/16

	Empty model		Model 1.1		Model 1.2		Model 2.1		Model 2.2	
	$\beta_{0ij} = 4.834 (0.057)$		$\beta_{0ij} = 2.628 (0.298)$		$\beta_{0ij} = 2.663 (0.279)$		$\beta_{0ij} = 3.254 (0.843)$		$\beta_{0ij} = 3.154 (0.829)$	
	Co-efficient $\beta$	p	Co-efficient $\beta$ (SE)	p	Co-efficient $\beta$ (SE)	p	Co-efficient $\beta$ (SE)	p	Co-efficient $\beta$ (SE)	p
<b>Class</b>	0.134 (0.033)	<.001	0.130 (0.034)	<.001	0.159 (0.038)	<.001	0.136 (0.035)	<.001	0.167 (0.040)	<.001
Procedural Fidelity and Quality (compared to low)			0.336 (0.196) – if mod	.048	0.129 (0.193) – if mod	.247	0.010 (0.006)	.051	0.005 (0.006)	.206
Participant responsiveness (compared to low)			0.428 (0.228) - if high	.035	0.227 (0.226) - if high	.162				
Participant reach (compared to low)			0.156 (0.167) – if mod	.180	0.307 (0.173) – if mod	.042	0.001 (0.004)	.402	0.002 (0.003)	.253
Dosage (compared to low)			0.067 (0.220) – if high	.383	0.181 (0.226) – if high	.214				
			-0.330 (0.210) – if mod	.081	-0.235 (0.204) – if mod	.128	-0.013 (0.008)	.055	-0.009 (0.008)	.134
			-0.249 (0.159) – if high	.063	-0.144 (0.157) – if high	.182				
			-0.259 (0.169) – if mod	.068	-0.251 (0.182) – if mod	.088	0.000 (0.000)	.500	0.000 (0.000)	.500
			-0.444 (0.216) – if high	.024	-0.464 (0.207) - if high	.015				
<b>Pupil</b>	0.701 (0.030)	<.001	0.555 (0.026)	<.001	0.528 (0.022)	<.001	0.555 (0.026)	<.001	0.530 (0.024)	<.001
Gender (if male)			-0.215 (0.050)	<.001	-0.215 (0.047)	<.001	-0.214 (0.050)	<.001	-0.211 (0.047)	<.001
FSM (if eligible)			-0.128 (0.059)	.015	-0.127 (0.054)	.009	-0.128 (0.059)	.015	-0.135 (0.053)	.005
T1 score			0.484 (0.033)	<.001	0.473 (0.030)	<.001	0.485 (0.033)	<.001	0.477 (0.030)	<.001
-2*Loglikelihood =	2546.831		2154.817				2156.049			
			X <sup>2</sup> (df=15, n=924) = 702.014, p <.001				X <sup>2</sup> (df=7, n=924) = 390.782, p <.001			

Table A22: Reading 2016/17

	Empty model		Model 1.1		Model 1.2		Model 2.1		Model 2.2	
	$\beta_{0ij} = 31.525 (0.680)$		$\beta_{0ij} = -1.647 (3.761)$		$\beta_{0ij} = -1.310 (3.812)$		$\beta_{0ij} = -12.031 (7.768)$		$\beta_{0ij} = -11.214 (7.809)$	
	Co-efficient $\beta$	p	Co-efficient $\beta$ (SE)	p	Co-efficient $\beta$ (SE)	p	Co-efficient $\beta$ (SE)	p	Co-efficient $\beta$ (SE)	p
<b>Class</b>	16.612 (4.423)	<.001	5.467 (1.622)	<.001	6.026 (1.684)	<.001	6.179 (1.777)	<.001	6.670 (1.851)	<.001
Fidelity and Quality (compared to low)			-0.164 (1.279) – if mod	.449	-0.311 (1.277) – if mod	.406	0.047 (0.044)	.146	0.050 (0.043)	.123
Participant responsiveness (compared to low)			2.583 (1.763) - if high	.075	2.473 (1.802) – if high	.089				
Participant reach (compared to low)			2.232 (1.046) – if mod	.020	1.925 (1.062) – if mod	.039	0.034 (0.029)	.124	0.020 (0.030)	.252
Dosage (compared to low)			1.319 (1.362) – if high	.164	1.196 (1.400) - if high	.200				
			0.451 (1.476) – if mod	.379	0.501 (1.533) - if mod	.372	0.075 (0.073)	.155	0.081 (0.074)	.138
			1.407 (1.530) – if high	.182	1.204 (1.536) – if high	.220				
			0.323 (3.206) – if mod	.460	0.594 (3.311) – if mod	.429	0.001 (0.001)	.500	0.001 (0.001)	.161
			0.407 (3.495) – if high	.453	0.758 (3.638) – if high	.417				
<b>Pupil</b>	90.557 (4.236)	<.001	38.818 (1.898)	<.001	39.590 (1.920)	<.001	38.795 (1.897)	<.001	40.020 (1.905)	<.001
Gender (if male)			-0.298 (0.433)	.245	-0.506 (0.414)	.111	-0.311 (0.433)	.236	-0.508 (0.441)	.125
FSM (if eligible)			-0.372 (0.509)	.233	-0.266 (0.469)	.284	-0.352 (0.509)	.245	-0.228 (0.495)	.323
T1 score			1.970 (0.059)	<.001	1.955 (0.058)	<.001	1.964 (0.059)	<.001	1.924 (0.059)	<.001
-2*Loglikelihood =	7121.366		5760.844				5764.169			
			X <sup>2</sup> (df=15, n=878) = 1360.522, p <.001				X <sup>2</sup> (df=7, n=878) = 1357.197, p <.001			

**Table A23: Concentration problems 2016/17**

	Empty model		Model 1.1		Model 1.2		Model 2.1		Model 2.2	
	$\beta_{0ij} = 2.539 (0.067)$		$\beta_{0ij} = -0.009 (0.559)$		$\beta_{0ij} = 0.100 (0.591)$		$\beta_{0ij} = 1.788 (1.269)$		$\beta_{0ij} = 2.014 (1.294)$	
	Co-efficient $\beta$	p	Co-efficient $\beta$ (SE)	p	Co-efficient $\beta$ (SE)	p	Co-efficient $\beta$ (SE)	p	Co-efficient $\beta$ (SE)	p
<b>Class</b>	0.144 (0.043)	<.001	0.144 (0.039)	<.001	0.168 (0.043)	<.001	0.184 (0.048)	<.001	0.195 (0.047)	<.001
Fidelity and Quality (compared to low)			0.344 (0.298) – if mod	.130	0.294 (0.204) – if mod	.079	0.005 (0.007)	.241	-0.001 (0.007)	.445
Participant responsiveness (compared to low)			0.081 (0.272) - if high	.383	0.059 (0.282) – if high	.417				
Participant reach (compared to low)			-0.315 (0.162) – if mod	.030	-0.354 (0.170) – if mod	.022	0.007 (0.005)	.085	-0.004 (0.005)	.214
Dosage (compared to low)			-0.343 (0.211) – if high	.057	-0.344 (0.213) - if high	.057				
			0.005 (0.232) – if mod	.492	-0.026 (0.248) - if mod	.460	-0.009 (0.012)	.229	-0.009 (0.012)	.228
			-0.150 (0.237) – if high	.267	-0.152 (0.241) – if high	.266				
			0.799 (0.483) – if mod	.055	0.795 (0.508) – if mod	.063	-0.000 (0.000)	.500	-0.000 (0.000)	.500
			0.448 (0.528) – if high	.182	0.485 (0.552) – if high	.192				
<b>Pupil</b>	1.165 (0.055)	<.001	0.626 (0.031)	<.001	0.622 (0.029)	<.001	0.626 (0.031)	<.001	0.638 (0.030)	<.001
Gender (if male)			0.231 (0.058)	<.001	0.226 (0.054)	<.001	0.232 (0.058)	<.001	0.223 (0.054)	<.001
FSM (if eligible)			0.132 (0.066)	.023	0.146 (0.062)	.009	0.132 (0.066)	.023	0.148 (0.066)	.013
T1 score			0.647 (0.027)	<.001	0.634 (0.025)	<.001	0.647 (0.028)	<.001	0.632 (0.026)	<.001
-2*Loglikelihood =	2900.520		2072.274				2080.557			
			X <sup>2</sup> (df=15, n=845) =828.246, p <.001				X <sup>2</sup> (df=7, n=845) = 819.963, p <.001			

Table A24: Disruptive behaviour 2016/17

	Empty model		Model 1.1		Model 1.2		Model 2.1		Model 2.2	
	$\beta_{0ij} = 1.737 (0.047)$		$\beta_{0ij} = -0.107 (0.430)$		$\beta_{0ij} = -0.064 (0.437)$		$\beta_{0ij} = 0.404 (0.972)$		$\beta_{0ij} = 0.779 (0.944)$	
	Co-efficient $\beta$	p	Co-efficient $\beta$ (SE)	p	Co-efficient $\beta$ (SE)	p	Co-efficient $\beta$ (SE)	p	Co-efficient $\beta$ (SE)	p
<b>Class</b>	0.066 (0.021)	<.001	0.085 (0.023)	<.001	0.089 (0.023)	<.001	0.106 (0.028)	<.001	0.106 (0.027)	<.001
Fidelity and Quality (compared to low)			0.310 (0.152) – if mod	.025	0.279 (0.154) – if mod	.039	0.009 (0.005)	.040	0.005 (0.005)	.161
Participant responsiveness (compared to low)			0.121 (0.209) – if high	.283	0.081 (0.210) – if high	.349				
Participant reach (compared to low)			-0.284 (0.124) – if mod	.015	-0.242 (0.121) – if mod	.026	-0.007 (0.004)	.044	-0.004 (0.003)	.095
Dosage (compared to low)			-0.359 (0.163) – if high	.018	-0.290 (0.161) – if high	.040				
			0.206 (0.179) – if mod	.130	0.188 (0.187) – if mod	.162	-0.002 (0.009)	.413	-0.005 (0.009)	.289
			0.054 (0.183) – if high	.383	0.002 (0.184) – if high	.496				
			0.315 (0.372) – if mod	.201	0.351 (0.378) – if mod	.179	-0.000 (0.000)	.500	-0.000 (0.000)	.500
			0.334 (0.407) – if high	.209	0.405 (0.413) – if high	.166				
<b>Pupil</b>	0.691 (0.032)	<.001	0.387 (0.019)	<.001	0.376 (0.019)	<.001	0.387 (0.019)	<.001	0.377 (0.018)	<.001
Gender (if male)			0.142 (0.045)	<.001	0.139 (0.041)	<.001	0.143 (0.045)	<.001	0.144 (0.040)	<.001
FSM (if eligible)			0.092 (0.051)	.036	0.107 (0.045)	.009	0.093 (0.051)	.035	0.099 (0.049)	.022
T1 score			0.810 (0.034)	<.001	0.765 (0.030)	<.001	0.812 (0.034)	<.001	0.779 (0.032)	<.001
-2*Loglikelihood =	2396.127		1663.663				1671.006			
			$X^2$ (df=15, n=845) =732.464, p <.001				$X^2$ (df=7, n=845) = 725.121, p <.001			



Table A25: Pro-social behaviour 2016/17

	Empty model		Model 1.1		Model 1.2		Model 2.1		Model 2.2	
	$\beta_{0ij} = 4.878 (0.062)$		$\beta_{0ij} = 3.472 (0.610)$		$\beta_{0ij} = 3.382 (0.595)$		$\beta_{0ij} = 3.612 (1.287)$		$\beta_{0ij} = 3.348 (1.233)$	
	Co-efficient $\beta$	p	Co-efficient $\beta$ (SE)	p	Co-efficient $\beta$ (SE)	p	Co-efficient $\beta$ (SE)	p	Co-efficient $\beta$ (SE)	p
<b>Class</b>	0.137 (0.036)	<.001	0.172 (0.044)	<.001	0.167 (0.040)	<.001	0.195 (0.049)	<.001	0.187 (0.044)	<.001
Fidelity and Quality (compared to low)			-0.485 (0.209) – if mod	.014	-0.377 (0.196) – if mod	.031	-0.013 (0.007)	.035	-0.010 (0.007)	.079
Participant responsiveness (compared to low)			-0.343 (0.288) - if high	.122	-0.213 (0.272) – if high	.220				
Participant reach (compared to low)			0.257 (0.171) – if mod	.072	0.232 (0.163) – if mod	.082	0.007 (0.005)	.085	0.004 (0.004)	.161
Dosage (compared to low)			0.348 (0.223) – if high	.065	0.245 (0.207) – if high	.071				
			-0.249 (0.245) – if mod	.158	-0.218 (0.241) – if mod	.187	-0.008 (0.012)	.253	-0.006 (0.012)	.310
			-0.355 (0.251) – if high	.085	-0.242 (0.235) - if high	.154				
			-0.259 (0.508) – if mod	.307	-0.576 (0.497) – if mod	.126	0.000 (0.000)	.500	0.000 (0.000)	.500
			-0.243 (0.556) – if high	.332	-0.350 (0.541) – if high	.260				
<b>Pupil</b>			0.504 (0.025)	<.001	0.500 (0.024)	<.001	0.504 (0.025)	<.001	0.501 (0.024)	<.001
Gender (if male)	0.675 (0.032)	<.001	-0.191 (0.051)	<.001	-0.187 (0.048)	<.001	-0.191 (0.051)	<.001	-0.186 (0.048)	<.001
FSM (if eligible)			-0.107 (0.059)	.036	-0.076 (0.053)	.077	-0.108 (0.059)	.034	-0.074 (0.058)	.101
T1 score			0.496 (0.034)	<.001	0.502 (0.032)	<.001	0.497 (0.034)	<.001	0.509 (0.030)	<.001
-2*Loglikelihood =	2400.530		1902.590				1907.215			
			X <sup>2</sup> (df=15, n=845) = 497.640, p <.001				X <sup>2</sup> (df=7, n=845) = 493.315, p <.001			

## Appendix 12: Security classification of trial findings

Rating	Criteria for rating			Initial score	Adjust	Final score
	Design	Power	Attrition			
5	Well conducted experimental design with appropriate analysis	MDES < 0.2	0-10%			
4	Fair and clear quasi-experimental design for comparison (e.g. RDD) with appropriate analysis, or experimental design with minor concerns about validity	MDES < 0.3	11-20%	4	Adjustment for Balance [ ]	4
3	Well-matched comparison (using propensity score matching, or similar) or experimental design with moderate concerns about validity	MDES < 0.4	21-30%		Adjustment for threats to internal validity [ ]	
2	Weakly matched comparison or experimental design with major flaws	MDES < 0.5	31-40%			
1	Comparison group with poor or no matching (E.g. volunteer versus others)	MDES < 0.6	41-50%			
0	No comparator	MDES > 0.6	over 50%			

- **Initial padlock score:** lowest of the three ratings for design, power and attrition = [4] padlocks
- **Reason for adjustment for balance** (if made): No adjustment necessary
- **Reason for adjustment for threats to validity:** No adjustment necessary
- **Final padlock score:** 4 padlocks

## Appendix 13: EEF cost rating

Cost ratings are based on the approximate cost per pupil per year of implementing the intervention over three years. More information about the EEF's approach to cost evaluation can be found [here](#). Cost ratings are awarded as follows:

Cost rating	Description
£ £ £ £ £	<i>Very low:</i> less than £80 per pupil per year.
£ £ £ £ £	<i>Low:</i> up to about £200 per pupil per year.
£ £ £ £ £	<i>Moderate:</i> up to about £700 per pupil per year.
£ £ £ £ £	<i>High:</i> up to £1,200 per pupil per year.
£ £ £ £ £	<i>Very high:</i> over £1,200 per pupil per year.

You may re-use this document/publication (not including logos) free of charge in any format or medium, under the terms of the Open Government Licence v3.0.

**OGL** This information is licensed under the Open Government Licence v3.0. To view this licence, visit <http://www.nationalarchives.gov.uk/doc/open-government-licence/>

Where we have identified any third party copyright information you will need to obtain permission from the copyright holders concerned. The views expressed in this report are the authors' and do not necessarily reflect those of the Department for Education.

This document is available for download at [www.educationendowmentfoundation.org.uk](http://www.educationendowmentfoundation.org.uk)



Education  
Endowment  
Foundation

The Education Endowment Foundation  
9th Floor, Millbank Tower  
21–24 Millbank  
London  
SW1P 4QP  
[www.educationendowmentfoundation.org.uk](http://www.educationendowmentfoundation.org.uk)