LJMU Research Online

Humphrey, N, Hennessey, A, Troncoso, P, Panayiotou, M, Black, L, Petersen, K, Wo, L, Mason, C, Ashworth, E, Frearson, K, Boehnke, JR, Pockett, RD, Lowin, J, Foxcroft, D, Wigelsworth, M and Lendrum, A

The Good Behaviour Game intervention to improve behavioural and other outcomes for children aged 7–8 years: a cluster RCT

http://researchonline.ljmu.ac.uk/id/eprint/16857/

Article

Public Health Research

The Good Behaviour Game intervention to improve behavioural and other outcomes for children aged 7–8 years: a cluster RCT

Neil Humphrey, Alexandra Hennessey, Patricio Troncoso, Margarita Panayiotou, Louise Black, Kimberly Petersen, Lawrence Wo, Carla Mason, Emma Ashworth, Kirsty Frearson, Jan R Boehnke, Rhys D Pockett, Julia Lowin, David Foxcroft, Michael Wigelsworth and Ann Lendrum

The Good Behaviour Game intervention to improve behavioural and other outcomes for children aged 7–8 years: a cluster RCT

Neil Humphrey[1]* Alexandra Hennessey,[1]
Patricio Troncoso,[1,2] Margarita Panayiotou,[1]
Louise Black,[1] Kimberly Petersen,[1] Lawrence Wo,[1]
Carla Mason,[1] Emma Ashworth,[3] Kirsty Frearson,[1]
Jan R Boehnke,[4] Rhys D Pockett,[5] Julia Lowin,[5]
David Foxcroft,[6] Michael Wigelsworth[1]
and Ann Lendrum[1]

[1]Manchester Institute of Education, University of Manchester, Manchester, UK
[2]Institute for Social Policy, Housing, Equalities Research, Heriot-Watt University, Edinburgh, UK
[3]School of Psychology, Liverpool John Moores University, Liverpool, UK
[4]School of Health Sciences, University of Dundee, Dundee, UK
[5]Swansea Centre for Health Economics, University of Swansea, Swansea, UK
[6]Department of Psychology, Health and Professional Development, Oxford Brookes University, Oxford, UK

*Corresponding author

This report should be referenced as follows:

Humphrey N, Hennessey A, Troncoso P, Panayiotou M, Black L, Petersen K, *et al*. The Good Behaviour Game intervention to improve behavioural and other outcomes for children aged 7–8 years: a cluster RCT. *Public Health Res* 2022;**10**(7).

Public Health Research

> **Criteria for inclusion in the *Public Health Research* journal**
> Reports are published in *Public Health Research* (PHR) if (1) they have resulted from work for the PHR programme, and (2) they are of a sufficiently high scientific quality as assessed by the reviewers and editors.
>
> Reviews in *Public Health Research* are termed 'systematic' when the account of the search appraisal and synthesis methods (to minimise biases and random errors) would, in theory, permit the replication of the review by others.

PHR programme

The Public Health Research (PHR) programme, part of the National Institute for Health and Care Research (NIHR), is the leading UK funder of public health research, evaluating public health interventions, providing new knowledge on the benefits, costs, acceptability and wider impacts of non-NHS interventions intended to improve the health of the public and reduce inequalities in health. The scope of the programme is multi-disciplinary and broad, covering a range of interventions that improve public health.

For more information about the PHR programme please visit the website: https://www.nihr.ac.uk/explore-nihr/funding-programmes/public-health-research.htm

This report

Abstract

The Good Behaviour Game intervention to improve behavioural and other outcomes for children aged 7–8 years: a cluster RCT

Neil Humphrey,[1]* Alexandra Hennessey,[1] Patricio Troncoso,[1,2] Margarita Panayiotou,[1] Louise Black,[1] Kimberly Petersen,[1] Lawrence Wo,[1] Carla Mason,[1] Emma Ashworth,[3] Kirsty Frearson,[1] Jan R Boehnke,[4] Rhys D Pockett,[5] Julia Lowin,[5] David Foxcroft,[6] Michael Wigelsworth[1] and Ann Lendrum[1]

[1]Manchester Institute of Education, University of Manchester, Manchester, UK
[2]Institute for Social Policy, Housing, Equalities Research, Heriot-Watt University, Edinburgh, UK
[3]School of Psychology, Liverpool John Moores University, Liverpool, UK
[4]School of Health Sciences, University of Dundee, Dundee, UK
[5]Swansea Centre for Health Economics, University of Swansea, Swansea, UK
[6]Department of Psychology, Health and Professional Development, Oxford Brookes University, Oxford, UK

*Corresponding author neil.humphrey@manchester.ac.uk

Background: Universal, school-based behaviour management interventions can produce meaningful improvements in children's behaviour and other outcomes. However, the UK evidence base for these remains limited.

Objective: The objective of this trial was to investigate the impact, value for money and longer-term outcomes of the Good Behaviour Game. Study hypotheses centred on immediate impact (hypothesis 1); subgroup effects (at-risk boys, hypothesis 2); implementation effects (dosage, hypothesis 3); maintenance/ sleeper effects (12- and 24-month post-intervention follow-ups, hypothesis 4); the temporal association between mental health and academic attainment (hypothesis 5); and the health economic impact of the Good Behaviour Game (hypothesis 6).

Design: This was a two-group, parallel, cluster-randomised controlled trial. Primary schools ($n = 77$) were randomly assigned to implement the Good Behaviour Game for 2 years or continue their usual practice, after which there was a 2-year follow-up period.

Setting: The trial was set in primary schools across 23 local authorities in England.

Participants: Participants were children ($n = 3084$) aged 7–8 years attending participating schools.

Intervention: The Good Behaviour Game is a universal behaviour management intervention. Its core components are classroom rules, team membership, monitoring behaviour and positive reinforcement. It is played alongside a normal classroom activity for a set time, during which children work in teams to win the game to access the agreed rewards. The Good Behaviour Game is a manualised intervention delivered by teachers who receive initial training and ongoing coaching.

Main outcome measures: The measures were conduct problems (primary outcome; teacher-rated Strengths and Difficulties Questionnaire scores); emotional symptoms (teacher-rated Strengths and Difficulties Questionnaire scores); psychological well-being, peer and social support, bullying

(i.e. social acceptance) and school environment (self-report Kidscreen survey results); and school absence and exclusion from school (measured using National Pupil Database records). Measures of academic attainment (reading, standardised tests), disruptive behaviour, concentration problems and prosocial behaviour (Teacher Observation of Child Adaptation Checklist scores) were also collected during the 2-year follow-up period.

Results: There was no evidence that the Good Behaviour Game improved any outcomes (hypothesis 1). The only significant subgroup moderator effect identified was contrary to expectations: at-risk boys in Good Behaviour Game schools reported higher rates of bullying (hypothesis 2). The moderating effect of the amount of time spent playing the Good Behaviour Game was unclear; in the context of both moderate ($\geq$ 1030 minutes over 2 years) and high ($\geq$ 1348 minutes over 2 years) intervention compliance, there were significant reductions in children's psychological well-being, but also significant reductions in their school absence (hypothesis 3). The only medium-term intervention effect was for peer and social support at 24 months, but this was in a negative direction (hypothesis 4). After disaggregating within- and between-individual effects, we found no temporal within-individual associations between children's mental health and their academic attainment (hypothesis 5). Last, our cost–consequences analysis indicated that the Good Behaviour Game does not provide value for money (hypothesis 6).

Limitations: Limitations included the post-test-only design for several secondary outcomes; suboptimal implementation dosage (mitigated by complier-average causal effect estimation); and moderate child-level attrition (18.5% for the primary outcome analysis), particularly in the post-trial follow-up period (mitigated by the use of full information maximum likelihood procedures).

Future work: Questions remain regarding programme differentiation (e.g. how distinct is the Good Behaviour Game from existing behaviour management practices, and does this makes a difference in terms of its impact?) and if the Good Behaviour Game is impactful when combined with a complementary preventative intervention (as has been the case in several earlier trials).

Conclusion: The Good Behaviour Game cannot be recommended based on the findings reported here.

Trial registration: This trial is registered as ISRCTN64152096.

Contents

CONTENTS

List of tables

List of figures

List of supplementary material

Report Supplementary Material 1 Syntax for models not specified in the SAP by hypothesis

Supplementary material can be found on the NIHR Journals Library report page (https://doi.org/10.3310/VKOF7695).

Supplementary material has been provided by the authors to support the report and any files provided at submission will have been seen by peer reviewers, but not extensively reviewed. Any supplementary material provided at a later stage in the process may not have been peer reviewed.

List of abbreviations

AIR	American Institutes for Research		NPD	National Pupil Database
CACE	complier-average causal effect		ONS	Office for National Statistics
CCA	cost–consequences analysis		PAX	PAXIS Institute
CFA	confirmatory factor analysis		PPI	patient and public involvement
CFI	comparative fit index		RCT	randomised controlled trial
CI	confidence interval		RI-CLPM	random-intercept cross-lagged panel model
CLPM	cross-lagged panel model		RMSEA	root-mean-square error of approximation
CONSORT	Consolidated Standards Of Reporting Trials		SAP	statistical analysis plan
EAL	English as an additional language		SD	standard deviation
EEF	Education Endowment Foundation		SDQ	Strengths and Difficulties Questionnaire
ES	effect size		SE	standard error
FIML	full information maximum likelihood		SEM	structural equation model
FSM	free school meal		SEND	special educational needs and disabilities
GBG	Good Behaviour Game		SRS	secure research service
GLLAMM	generalised linear latent and mixed model		T1	time 1
HGRT	Hodder Group Reading Test		T2	time 2
IPE	implementation and process evaluation		T3	time 3
			T4	time 4
IRR	incidence rate ratio		T5	time 5
ITT	intention to treat		TLI	Tucker–Lewis Index
KS	key stage		TOCA-C	Teacher Observation of Child Adaptation Checklist
LCSFT	life course/social field theory			
MoA	memorandum of agreement		TSC	Trial Steering Committee
NICE	National Institute for Health and Care Excellence		WLSMV	weighted least squares means and variance adjusted
NIHR	National Institute for Health and Care Research			

Plain English summary

Up to 1 hour per week of learning time in primary schools is lost because of low-level disruptive behaviour (e.g. calling out and fidgeting). The Good Behaviour Game is an approach used by teachers to improve behaviour in the classroom. It includes classroom rules, teamwork and positive reinforcement of good behaviour.

We asked:

1. What is the impact of the Good Behaviour Game on children's behaviour problems and other outcomes (e.g. well-being)?
2. Is the Good Behaviour Game particularly effective for boys showing signs of behaviour problems?
3. Does the amount of time spent playing the Good Behaviour Game make a difference to children's outcomes?
4. Is there any impact 1 or 2 years after the intervention has finished?
5. Are children's mental health and their reading scores related over time?
6. Does the Good Behaviour Game provide value for money?

A total of 77 primary schools (> 3000 children) were allocated by chance to deliver (or not deliver) the Good Behaviour Game for 2 years. Data were collected a further 2 years after the intervention had ended.

We found:

1. There was no evidence that the Good Behaviour Game improved children's behaviour problems or other outcomes.
2. The Good Behaviour Game was not effective for boys showing signs of behaviour problems – in fact, it may have led to increased experiences of bullying for them.
3. The amount of time spent playing the Good Behaviour Game did not appear to influence children's outcomes, except that playing the game for longer reduced well-being and also reduced school absence.
4. We found little evidence of impact in follow-up assessments, except that children in Good Behaviour Game schools reported lower levels of peer and social support 2 years after the intervention had ended.
5. Children's mental health and their reading scores did not appear to be related over time.
6. The Good Behaviour Game did not provide value for money.

Based on these findings, we cannot recommend the Good Behaviour Game as a way to improve children's behaviour or other outcomes.

Scientific summary

Background

Children's behaviour in primary schools in England is mostly very good. Despite this, it is estimated that up to 1 hour of learning is lost each day as a consequence of low-level disruption in the classroom (e.g. fidgeting, calling out). Universal behaviour management interventions such as the Good Behaviour Game (GBG) aim to prevent disruptive behaviour in the classroom, with consequent improvements in a range of health- and education-related outcomes.

The GBG has an impressive international evidence base. There have been 14 randomised trials of the GBG, spanning seven countries. Among those that have reported findings at the intention-to-treat level, and for which the specific effects of the intervention can be isolated, most note significant effects on behavioural and other outcomes. The size of these effects is generally in line with those reported in meta-analytic studies of universal behaviour management interventions. However, there are some notable exceptions to this trend that report null results. Furthermore, relatively little is known about the medium- and long-term effects of the GBG, or the potential moderating role of implementation compliance.

The GBG is a promising intervention, but, prior to the current study, it had never been rigorously evaluated in England. We report findings from the first randomised controlled trial of the intervention in English primary schools, addressing a number of significant gaps in the evidence base.

Objectives

- To determine the impact of the GBG on health- and education-related outcomes for children.
- To determine the impact of the GBG on a variety of outcomes for boys at risk of developing conduct problems.
- To determine the extent to which the effects of the GBG vary as a function of intervention compliance (i.e. dosage).
- To determine whether or not the effects of the GBG are sustained (or emerge) over time.
- To assess the temporal association between mental health and academic attainment.
- To assess the health economic impact of the GBG.

Methods

A two-group, parallel, cluster-randomised controlled trial design was utilised, with schools as the unit of randomisation. Schools allocated to the intervention arm of the trial implemented the GBG throughout the school years 2015/16 and 2016/17. Those allocated to the usual-practice arm of the trial continued their existing approaches to managing behaviour during this period. The random allocation of schools was conducted independently of the authors by the Clinical Trials Unit at the Manchester Academic Health Science Centre (Manchester, UK), and, using minimisation, was balanced by school size and the proportion of children eligible for free school meals.

Intervention

The core components of the GBG are classroom rules, team membership, monitoring behaviour and positive reinforcement. In brief, children work in teams to win the game to access the agreed rewards.

The game is played alongside a normal classroom activity for a specified period of time, during which the teacher monitors infractions of four rules: we will (1) work quietly, (2) be polite to others, (3) only get out of our seats with permission and (4) follow directions. Teams with four or fewer infractions at the end of the game win and are rewarded. Over time, the GBG evolves in terms of the frequency and duration of play, and the nature and timing of rewards. Teachers implementing the GBG are supported by external coaches, who model game sessions, observe and provide feedback on implementation, offer ad hoc e-mail and telephone support, and provide additional/booster training or information sessions as required.

Participants

Participants were children ($n = 3084$) in Year 3 (aged 7–8 years) attending 77 participating primary schools (GBG, $n = 38$; usual practice, $n = 39$).

Outcome measures

The immediate post-intervention outcomes that we assessed were children's conduct problems [primary outcome: assessed using the teacher-rated Strengths and Difficulties Questionnaire (SDQ)], psychological well-being (assessed using the self-report Kidscreen survey), emotional symptoms (assessed using the teacher-rated SDQ), peer and social support (assessed using the self-report Kidscreen survey), school environment (assessed using the self-report Kidscreen survey), school absence (assessed using National Pupil Database records), bullying (i.e. social acceptance, assessed using the self-report Kidscreen survey) and exclusion from school (assessed using National Pupil Database records). Academic attainment (reading, assessed using standardised tests), disruptive behaviour, concentration problems and prosocial behaviour (assessed using the Teacher Observation of Child Adaptation Checklist) were also collected during the 2-year follow-up period.

The primary outcome was assessed at baseline, post intervention and at the 12- and 24-month follow-ups. Secondary outcome measures were assessed post intervention and at the 12- and 24-month follow-ups.

In addition, data on intervention compliance (i.e. dosage) were collected throughout the 2-year intervention period.

Results

There was no evidence that the GBG led to improvements in any of the above outcomes immediately after the intervention period (objective 1). The only significant subgroup moderator effect that was identified was contrary to expectations: at-risk boys in GBG schools reported higher rates of bullying at the end of the intervention period [effect size (ES) –0.563, 95% confidence interval (CI) –0.716 to –0.409; objective 2]. The evidence that intervention outcomes were moderated by the amount of time spent playing the GBG was minimal and somewhat conflicting; in the context of both moderate ($\geq$ 1030 minutes) and high ($\geq$ 1348 minutes) intervention compliance, there were significant negative effects on children's psychological well-being (moderate compliance, ES –0.241, 95% CI –0.312 to –0.170; high compliance, ES –0.294, 95% CI –0.365 to –0.223), but significant positive effects on school absence (moderate compliance, incidence rate ratio 0.519, 95% CI 0.450 to 0.598; high compliance, incidence rate ratio 0.510, 95% CI 0.371 to 0.701; objective 3). There was no evidence of the emergence of intervention effects at the 12-month or 24-month follow-ups on any outcomes, with the exception of a potentially negative effect on peer and social support (ES –0.195, 95% CI –0.265 to –0.125; objective 4). After disaggregating within- and between-individual effects, we found no temporal within-individual

associations between children's mental health and their academic attainment (objective 5). Last, our cost–consequences analysis indicated that the GBG does not provide value for money, with implementation costs of £275.68 per child, no attendant difference found in primary or secondary outcomes, and no difference in exclusion costs (objective 6).

Conclusions

On the basis of the findings reported here, it is not possible to recommend the GBG as a way to improve children's health- and education-related outcomes. However, we note that intervention compliance was suboptimal and, although our analyses indicated that outcomes mostly did not vary as a function of dosage, we cannot rule out the possibility that the minimum effective dose was not reached, even in our high-compliance settings. Nonetheless, the dosage reported was achieved in an efficacy trial context in which initial training and ongoing coaching support for teachers, subsidised intervention costs for schools, additional provision for data monitoring made available by our research team, and developer support for the delivery team were available. In other words, while we may have seen more evidence of meaningful intervention effects with significantly higher levels of implementation than were observed here, it is very unlikely that such levels would ever be achieved if the GBG were implemented at scale in England, in which case such a comprehensive implementation support system would be absent.

Other possible explanations for our results include cultural incompatibility and insufficient programme differentiation. In relation to the former, many teachers reported struggling with certain mandated intervention procedures, most notably not being able to directly interact or intervene with pupils during gameplay. With regard to the latter, our survey of teachers' behaviour management strategies revealed that those in the control arm of the trial were enacting practices that mirrored some of the core components of the GBG (e.g. classroom rules, team membership, monitoring behaviour and positive reinforcement). Given this, it is possible that the null results observed were due to the fact the intervention was insufficiently differentiated from the usual practice of schools.

The findings of this study raise a number of questions that future research might usefully seek to answer. Below, we outline some key gaps and provide an indication of what future studies might look like to address these:

- Who benefits from higher levels of dosage of interventions like the GBG?
  To address this question, future research should incorporate extensions of complier-average causal effect models (which account for implementation variability) to include subgroup moderator analyses (which facilitate the examination of differential gains among specified groups within a trial sample).
- Does the level of differentiation between the GBG and existing behaviour management practices in the classroom matter?
  To address this question, future research should examine whether the magnitude of intervention effects vary by level of programme differentiation. One might, for example, predict larger effects in 'high-differentiation' settings, where the constituent components of the GBG are novel, than in 'low-differentiation' settings in which they are less distinct from existing practice.
- Does the GBG have an impact if it is delivered in combination with another intervention(s)?
  To address this question, future research should use factorial trial designs, which enable the examination of an interaction between two or more interventions (e.g. control, GBG only, other intervention only, GBG and other intervention in combination).
- Do interventions like the GBG have an impact on the developmental process of growth?
  To address this question, future research should use growth curve models (as opposed to point-in-time estimates) that can examine the impact of interventions such as the GBG on developmental trajectories.

Public and patient involvement

The director of Common Room (Leeds, UK) and a team of six young research advisors undertook a range of activities throughout the study, including attendance at and contribution to Trial Steering Committee meetings; input and feedback on a range of study materials (e.g. child self-report surveys, standardised survey instructions, debriefs) and dissemination outputs [e.g. a short film on YouTube (YouTube, LLC, San Bruno, CA, USA) to present project findings in an accessible manner to non-academic audiences]; and focus groups in schools to discuss the experiences of children who had taken part in the GBG.

Trial registration

This trial is registered as ISRCTN64152096.

Funding

Chapter 1 Introduction

Parts of this chapter are reproduced or adapted with permission from Humphrey *et al.*[1] Contains information licensed under the Open Government Licence v3.0. URL: www.nationalarchives.gov.uk/doc/open-government-licence/version/3/.

Parts of this chapter are also reproduced or adapted with permission from the Good Behaviour Game (GBG) trial protocol [available from the National Institute for Health and Care Research (NIHR) project web page: www.journalslibrary.nihr.ac.uk/programmes/phr/145238].

Universal approaches to behaviour management in schools

Children's behaviour is considered to be good or outstanding in the overwhelming majority of primary schools in England.[2] Despite this, it is estimated that up to 1 hour of learning is lost each day as a consequence of low-level disruption in the classroom. Teachers surveyed by the Office for Standards in Education, Children's Services and Skills (Ofsted) identified talking, calling out and fidgeting as key problems.[3] In addition, for a minority of children, more serious concerns about behaviour are evident: approximately 5% of children in primary school display the aggressive, defiant and antisocial behaviours that characterise behavioural (or conduct) disorders, with prevalence rates more than twice as high among boys (6.7%) as among girls (3.2%).[4] In the short term, such difficulties create significant challenges for behaviour management and erode children's academic development.[5] In the longer term, childhood conduct problems, particularly among boys, are associated with a twofold to threefold increase in early adulthood public-sector costs (mainly via the criminal justice system) and significantly higher rates of unemployment.[6,7] Accordingly, developing the evidence base regarding the most effective behaviour management strategies has been set as a research priority by both the government[8] and National Institute for Health and Care Excellence (NICE).[9]

Universal approaches to behaviour management in schools can be usefully classified according to their focus: teachers' behaviour, teacher–student relationships, students' behaviour and/or students' social–emotional development. Evidence from a recent meta-analysis indicates that such approaches can produce meaningful improvements in children's behavioural ($g = 0.24$), academic ($g = 0.17$), social–emotional ($g = 0.21$) and other ($g = 0.26$) outcomes.[10] The GBG is an example of a behaviour management approach that focuses primarily on students' behaviour. More specifically, it is an 'interdependent group-oriented contingency management procedure'[11] whose core components are classroom rules, team membership, monitoring behaviour and positive reinforcement. In brief, children work in teams to win the GBG to access the agreed rewards. It is played alongside a normal classroom activity for a specified period of time, during which the teacher monitors adherence to four rules: we will (1) work quietly, (2) be polite to others, (3) only get out of our seats with permission and (4) follow directions. Teams who break these rules four times or fewer win the game and are rewarded.[12] Over time, the GBG evolves in terms of the frequency and duration of play, and the nature and timing of rewards. It is underpinned by behaviourism (e.g. contingency management and the reproduction of rewarded behaviour),[13] social learning theory (e.g. learning of appropriate behaviour modelled effectively by other team members)[14] and life course/social field theory (LCSFT) (e.g. promotion of adaptive processes to enable children to meet social task demands in the classroom).[15] A more comprehensive and detailed description of the intervention is provided in *Chapter 2*.

It is important to note from the outset that the GBG has evolved since the first report on the intervention was published over 50 years ago. Indeed, there are multiple versions evident in the literature, including (but not limited to) the American Institutes for Research (AIR) model used in this trial, the PAXIS Institute (PAX) model (known as PAX GBG), and various cultural and other adaptations that have been developed as the intervention has been implemented in different contexts over time. Although they all share common

core components (e.g. classroom rules, team membership, monitoring behaviour and positive reinforcement), each version of the GBG is also somewhat distinct. For example, the other most widely used version, PAX GBG, differs from the AIR GBG model in terms of (a) the language used to describe rule adherence and rule breaks (referred to as 'PAX' and 'spleems', respectively), (b) the game reward threshold (teams with 3 or fewer spleems, as opposed to 4 or fewer rule breaks, access the agreed reward), (c) use of parent activities to promote generalisation of self-regulation skills to the home environment and (d) various additional 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).[16] In terms of the evidence base discussed below, it is not always abundantly clear exactly which version of the GBG has been trialled. Although randomised controlled trials (RCTs) of PAX GBG are always clearly labelled as such,[17–22] others do not clearly articulate the underpinning model. In correspondence with AIR and PAX, we have therefore attempted to provide some clarity as part of our summary of the available evidence (*Table 1*).

Evidence base for the Good Behaviour Game

There have been 14 RCTs of the GBG to date: six in the USA,[17,18,23,29,33,37] two each in the Netherlands[31,32] and Canada;[19,35] and one each in Belgium,[34] Northern Ireland,[36] Estonia[21] and England.[1] These trials represent the standard context for implementation of the GBG (e.g. whole class delivery during a normal school day); a fifteenth RCT based in the USA reports on the impact of the intervention in the context of an after-school programme.[22] *Table 1* provides a summary of the designs and findings of these studies.

To aid interpretation of *Table 1*, we note that (1) the Turkkan version of the GBG (manualised by Jaylan Turkkan at Johns Hopkins University), referenced in relation to two trials[17,23] is a precursor to the AIR model that uses the same procedures and rules (i.e. four or fewer infractions to win the game); (2) that the two follow-up analyses for one trial[23] have conflicting findings, with one reporting null effects[25] and the other reporting significant intervention effects for the aggressive male subgroup;[24] and (3) that a definitive sample size is not provided for one trial,[18] but the authors report that 8–10 students were sampled randomly for each of 188 teachers.

Inspection of *Table 1* reveals a number of trends in study design that are directly pertinent to the current study.

First, several RCTs have trialled the GBG in combination with other interventions, often using designs that mean that the effects of each cannot be properly isolated.[17,18,35,38]

Second, intention-to-treat (ITT) findings, in which analyses include every participant according to their randomisation, irrespective of their characteristics (e.g. sex, baseline risk status) and/or post-randomisation events (e.g. non-compliance, withdrawal),[39] have been somewhat variable. Some trials[17,23,33] have failed to report true ITT findings, raising the risk of bias in these studies. Among those that have reported ITT findings, and in which the specific effects of the GBG can be isolated, most note significant intervention effects on behavioural and other outcomes,[19,21,31,32,34,36] although there are a couple of notable exceptions,[1,37] including the trial on which the current study builds (see *The Good Behaviour Game in England*).

Third, medium- (i.e. 12–24 months) and long-term (i.e. > 24 months) post-intervention follow-up is rare: only four trials[17,23,29] – including the current study – have included any kind of follow-up beyond the immediate post test, precluding assessment of sleeper and/or maintenance effects in most cases.

Fourth, the reporting of implementation data is highly variable. In many cases, either it is not reported at all[23,32,33] or the data are extremely sparse.[17,34,38] Furthermore, for an intervention explicitly premised on the frequency and duration of delivery, the reporting of dosage data is surprisingly absent from many trials, precluding rigorous analysis of the moderating effects of intervention compliance. Where dosage is documented, it is typically self-reported by teachers, a method that is known to be subject to bias and impression management.[40]

TABLE 1 Randomised trials of the GBG

Authors	Year	Country	GBG version	RCT arms	Cluster unit (n)	Sample size (n)	Sample age (at baseline) (years)	Intervention duration	Level of implementation	Child outcomes (source/informant)	Intention-to-treat findings	Subgroup analysis	Subgroup findings	Longer-term follow-up	Longer-term findings
Dolan et al.[23]	1993	USA	Turkkan/AIR	GBG, mastery learning, usual practice	Classrooms (42)	864	6–7	6 months	N/A	Aggressive behaviour (teacher rating, peer assessment), shy behaviour (teacher rating, peer assessment)	N/A	Sex, baseline level of outcomes	GBG reduced aggressive behaviour (teacher rating) of boys, girls, and those among those rated highest at baseline; reduced aggressive behaviour (peer nominated) of boys, girls, and those among those rated highest at baseline; reduced shy behaviour of boys and girls	Up to age 11–12 years: Kellam et al.[24,25] Up to age 19–21 years: Kellam et al.,[15-26] Wilcox et al.[27]	Up to age 11–12 years: the GBG reduced aggressive behaviour among boys rated highest at baseline; up to age 19–21 years: the GBG reduced suicidal ideation and attempts, drug abuse and smoking among males, alcohol abuse and antisocial personality disorder among males and females, and antisocial personality disorder among males rated highest in aggression at baseline
Ialongo et al.[17]	1999	USA	Turkkan/AIR	Combined GBG, curriculum enhancements, and targeted support; family–school partnership; usual practice	Schools (27)	678	5–7	1 year	Five out of nine GBG classrooms were classified as high fidelity and four were classed as low fidelity	Maths, reading (both standardised tests), attention/concentration (teacher and parent ratings), aggressive behaviour; shy behaviour (teacher, parent, and peer report)	N/A	Sex, baseline level of outcomes	GBG reduced aggressive behaviour of boys, particularly those rated highest at baseline; and improved maths and reading among boys rated lowest at baseline	Up to age 11–12 years: Ialongo et al.[28]	GBG reduced conduct problems, diagnostic criteria for conduct disorder, fixed-term exclusions and rates of child mental health service use
Reid et al.[29]	1999	USA	LIFT	Combined GBG, social skills and problem-solving curriculum, school–parent communication and parent training; usual practice	Schools (12)	671	6–7 (cohort 1) and 10–11 (cohort 2)	10 weeks	Reach: 90%	Physical aggression (independent observation), positive behaviour with peers (teacher rating)	GBG reduced physical aggression and improved positive behaviour with peers	N/A	N/A	Up to age 14 years (cohort 2): Eddy et al.[30]	GBG reduced onset of police arrest and patterned alcohol use

TABLE 1 Randomised trials of the GBG (continued)

Authors	Year	Country	GBG version	RCT arms	Cluster unit (n)	Sample size (n)	Sample age (at baseline) (years)	Intervention duration	Level of implementation	Child outcomes (source/informant)	Intention-to-treat findings	Subgroup analysis	Subgroup findings	Longer-term follow-up	Longer-term findings
van Lier et al.[31]	2004	The Netherlands	AIR (Dutch adaptation)	GBG, usual practice	Classrooms (31)	666	6–7 years	2 years	Nine out of 13 schools implemented the GBG 'completely'; three implemented the GBG but did not move to generalisation phase; and one implemented the GBG poorly	ADH, oppositional defiant disorder, conduct problems (teacher ratings)	GBG reduced ADH	ADH-latent classes	GBG reduced ADH, oppositional defiant disorder and conduct problems symptom trajectories of intermediate ADH class	N/A	N/A
Witvliet et al.[32]	2009	The Netherlands	AIR (Dutch adaptation)	GBG, usual practice	Classrooms (47)	758	7–9 years	2 years	N/A	Externalising behaviour (teacher rating), acceptance, mutual friendships, proximity to others (peer nomination)	GBG reduced externalising behaviour and improved acceptance, mutual friendships and proximity to others	Sex	GBG reduced externalising behaviour among boys	N/A	N/A
Hansen et al.[33]	2010	USA	All stars challenge	GBG, usual practice	Schools (11)	491	10–11	6 months	N/A	Physical aggression risk, social aggression risk, shyness risk, unawareness of social norms risk, overall risk (teacher rating)	N/A	Baseline risk rating (e.g. no risk, some risk, high risk)	GBG reduced shyness, increased awareness of social norms and reduced overall risk	N/A	N/A
Leflot et al.[34]	2010	Belgium	AIR (Dutch adaptation)	GBG, usual practice	Classrooms (30)	570	7–9	2 years	Fidelity: nine out of 12	Hyperactivity, oppositional behaviour (both peer nomination/ rating), on-task behaviour, talking out, and out-of-seat behaviour (independent observation)	GBG improved on task, and reduced talking out and oppositional behaviours	Sex	Null	N/A	N/A

Authors	Year	Country	GBG version	RCT arms	Cluster unit (n)	Sample size (n)	Sample age (at baseline) (years)	Intervention duration	Level of implementation	Child outcomes (source/informant)	Intention-to-treat findings	Subgroup analysis	Subgroup findings	Longer-term follow-up	Longer-term findings
Dion et al.[35]	2011	Canada	Attention le jis!	Combined GBG and peer tutoring, peer tutoring only, usual practice	Schools (30)	409	6–7	6 months	Fidelity: 95% Participant responsiveness: 90%	Attention (observation), reading (word recognition, non-word recognition and comprehension via standardised tests)	Combined GBG and peer tutoring improved attention	Baseline attention levels	Null	N/A	N/A
Humphrey et al.[1]	2018	England	AIR	GBG, usual practice	Schools (77)	3084	6–7	2 years	Fidelity: 69.95% Participant responsiveness: 71.79% Reach: 95.27% Dosage duration: 24.82 minutes per week Dosage frequency: 1.74 games per week Nine GBG schools ceased implementation prior to the end of the main trial	Reading (standardised tests), concentration problems, disruptive behaviour, prosocial behaviour (teacher rating)	Null	FSMs; boys at risk of conduct problems	Null	Up to age 10–11 years: the current study	This report (see *Chapter 3, Results*)

TABLE 1 Randomised trials of the GBG (*continued*)

Authors	Year	Country	GBG version	RCT arms	Cluster unit (*n*)	Sample size (*n*)	Sample age (at baseline) (years)	Intervention duration	Level of implementation	Child outcomes (source/informant)	Intention-to-treat findings	Subgroup analysis	Subgroup findings	Longer-term follow-up	Longer-term findings
Jiang *et al.*[19]	2018	Canada	PAX	GBG, usual practice	Schools (144)	3393	6–9	1 year	Dosage: played at least once per day in 70% of cases	Emotional symptoms, conduct problems, peer problems, ADH, prosocial behaviour (all teacher ratings)	GBG improved prosocial behaviour and reduced emotional symptoms, conduct problems, peer problems and ADH	Sex, socioeconomic status, baseline risk status	GBG reduced conduct problems among boys, improved prosocial behaviour and reduced peer problems among children with low socioeconomic status and improved all outcomes among high-risk children	N/A	N/A
O'Keeffe[36]	2019	Northern Ireland	PAX	GBG, usual practice	Schools (17)	353	6–8	12 weeks	Dosage frequency: games three times per day; Dosage duration: 20–60 minutes per day	Self-regulation, self-esteem (both child ratings); concentration, prosocial behaviour, disruptive behaviour, emotional symptoms, conduct problems, ADH, peer problems, prosocial behaviour, undesirable behaviours (all teacher ratings); co-operative learning (peer ratings)	GBG improved self-regulation	Sex, socioeconomic status, English as an additional language, special educational needs	GBG improved prosocial behaviour among boys, reduced disruptive behaviour and hyperactivity among children with special educational needs, reduced concentration problems and improved prosocial behaviour among children with low socioeconomic status, reduced prosocial behaviour among children with English as an additional language	N/A	N/A

Authors	Year	Country	GBG version	RCT arms	Cluster unit (n)	Sample size (n)	Sample age (at baseline) (years)	Intervention duration	Level of implementation	Child outcomes (source/informant)	Intention-to-treat findings	Subgroup analysis	Subgroup findings	Longer-term follow-up	Longer-term findings
Ialongo et al.[37]	2019	USA	PAX	Combined GBG and PATHS, GBG, usual practice	Schools (27)	5611	5–10	1 year	Combined GBG and PATHS: • Dosage frequency: 154 games • Dosage duration: 1583 minutes • Fidelity and quality score: 3/4 GBG only: • Dosage frequency: 150 games • Dosage duration: 1432 minutes • Fidelity and quality score: 3/4	Readiness to learn, social competence, emotion regulation, authority acceptance (teacher rating), total problem behaviour (observation)	Combined GBG and PATHS: reduced total problem behaviour GBG only: null	Sex, ethnicity, FSMs, grade level, cohort, baseline level of outcomes	Combined GBG and PATHS: • Improved readiness to learn, social competence, emotion regulation and authority acceptance among those rated lowest at baseline • Improved authority acceptance for those rated lowest and reduced total problem behaviour among those rated highest at baseline in grades K–2 • Improved readiness to learn and social competence among those rated lowest at baseline in Grades 3–5 • Improved emotion regulation and reduced total problem behaviour in Grades 3–5 GBG only: • Reduced total problem behaviour among those rated highest at baseline	N/A	N/A

TABLE 1 Randomised trials of the GBG (*continued*)

Authors	Year	Country	GBG version	RCT arms	Cluster unit (*n*)	Sample size (*n*)	Sample age (at baseline) (years)	Intervention duration	Level of implementation	Child outcomes (source/informant)	Intention-to-treat findings	Subgroup analysis	Subgroup findings	Longer-term follow-up	Longer-term findings
													• Improved authority acceptance for those rated lowest and reduced total problem behaviour among those rated highest at baseline in grades K–2		
Streimann *et al.*[21]	2020	Estonia	PAX	GBG, usual practice	Schools (46)	708	7–9	2 years	Fidelity: 25/30 Dosage frequency: 19/23 teachers reported playing the GBG daily	Total mental health difficulties, prosocial behaviour (teacher and parent ratings), ADH (parent ratings), classroom behaviour (teacher ratings)	GBG reduced total mental health difficulties and improved prosocial and classroom behaviour	Sex, baseline risk status	Null	N/A	N/A
Tolan *et al.*[18]	2020	USA	PAX	Combined GBG and My Teaching Partner, usual practice	Classrooms (188)	Circa 1692	5–9	1 year	Dosage frequency: games 12 times per week Dosage duration: 76.26 minutes per week Fidelity: 3/4	Socially disruptive behaviour, off-task behaviour, student compliance (all independent observation); reading and maths (both standardised tests)	Null	Baseline socially disruptive behaviour and teacher distress	GBG reduced socially disruptive and off-task behaviour in classrooms with high teacher distress at baseline, reduced socially disruptive and off-task behaviour and increased student compliance and maths in classrooms with high teacher distress and socially disruptive behaviour at baseline	N/A	N/A

ADH, attention deficit/hyperactivity; K–2, kindergarten to second grade; LIFT, Linking the Interests of Families and Teachers; N/A, not applicable; PATHS, Promoting Alternative Thinking Strategies.

Note
Shading denotes the Education Endowment Foundation (EEF)-funded GBG trial that the current report augments.

Fifth, analysis of subgroup effects is commonplace and typically focuses on sex and/or risk status, with the latter usually defined by elevated behaviour problems at baseline.

Last, most RCTs of the GBG have been modestly sized (e.g. < 50 clusters and < 1000 participants), placing limits on statistical power of analyses.

A recent meta-analysis[41] captured six of the 14 GBG trials. The two most recent US trials,[18,37] the most recent Canadian trial,[19] and the trials in Northern Ireland,[36] Estonia[21] and England[1] were concluded or published after the closing census date for the meta-analysis; two early US-based trials were also excluded for unknown reasons. The meta-analysis indicated that the intervention significantly outperformed comparison conditions in main effect/ITT analyses for three out of the six outcomes that were examined: teacher-rated conduct problems, peer-rated conduct problems and peer-rated peer relations (the other three outcomes were teacher-rated inattention, reading performance and teacher-rated peer relations).[41] The size of the intervention effect for these three outcomes ($g = 0.1$–0.2) was broadly in line with the above-noted findings of a meta-analysis of universal approaches to behaviour management.[10] Among the eight GBG RCTs not included in the meta-analysis,[1,18,19,21,29,33,36,37] three reported null ITT results,[1,18,37] and one did not report ITT findings.[33] The four that reported a significant main effect of the intervention reported a range of effect sizes (ESs), from 0.11 to 0.42, on behavioural and related outcomes.[19,21,29,36] Taken as a whole, these findings indicate small or moderate overall effects of the GBG when conventional ES thresholds are applied.[42] This probably reflects the fact that children's behaviour is typically very good in most schools,[2] with very few children displaying the symptoms of conduct or other problems at the outset of any given trial.[4]

Alongside the main effect estimates provided by ITT analysis, it is also important to consider moderated effects. Three treatment effect modifiers are particularly pertinent here: subgroups, implementation and timing of follow-up. First, it is widely recognised that children do not respond uniformly to exposure to universal interventions. Accordingly, subgroup analyses can be very informative, provided that they are specified in advance, are informed by theory and/or research, and include clear specification of the expected direction of effects and population subgroup(s) of interest (using features measured pre randomisation, e.g. demographic characteristics, individual differences at baseline and/or family factors).[43]

The effects of the GBG appear to vary by baseline risk status (e.g. higher levels of difficulties) and/or sex. In relation to baseline risk status, it stands to reason that those whose behaviour is already a significant cause for concern would stand to benefit the most from the GBG, especially given its emphasis on adaptive socialisation processes (e.g. alerting children to and rewarding them for meeting social task demands in the classroom). It is perhaps unsurprising that several GBG trials found amplified intervention effects among children considered 'at risk' because of their elevated levels of problematic behaviour.[17,23,33,37] In relation to sex, the intervention procedures may particularly appeal to boys, given the gendered socialisation of competitiveness.[44] A recent trial of the GBG in Canada found evidence to support this, with significantly greater reduction in conduct problems among boys than among girls.[19] The intersection of these two factors – that is, boys at risk of developing conduct problems – is a specific focus in this trial. Given the aforementioned early adulthood outcomes for this particular stratum of the population,[6,7] research that rigorously establishes efficacious, early, preventative strategies would be particularly welcome.[9] Promisingly, there is some existing evidence of amplified gains in this subgroup following exposure to the GBG.[24]

The second potential treatment effect moderator of note is variability in implementation. Such variability is considered to be inevitable, particularly in the case of universal school-based interventions,[45] and the accumulated evidence base suggests that it is associated with variability in the achievement of intended outcomes.[46] In the GBG, teachers may vary the frequency and/or duration of game sessions (i.e. dosage), their adherence to prescribed procedures (i.e. fidelity) and any associated changes to these (i.e. adaptations), and the extent to which they play the game in an enthusiastic and engaging manner (i.e. quality). Whether or not the game is played with all children in a given class (i.e. reach) and how they react when it is played (i.e. participant responsiveness) may also be important.

How similar or different the game is to existing behaviour management approaches in a given classroom (i.e. programme differentiation) and/or those against which it is being compared (i.e. control group activity) is also likely to contribute to its relative success. As noted above, there has been remarkably little empirical scrutiny of the extent to which variability in one or more of these implementation dimensions moderates treatment effects in RCTs of the GBG; instead, the norm has been to simply provide descriptive summaries.[41] A notable exception is Ialongo et al.'s trial,[17] in which the authors' per-protocol analysis (i.e. intervention schools divided into 'high' and 'low' implementation groups on the basis of fidelity scores) indicated an association between implementation level and the magnitude of certain intervention outcomes. However, caution is required in interpreting such findings given that per-protocol analysis compromises the randomised design.

Here, the application of complier-average causal effect (CACE) estimation (see *Chapter 2*) and related instrumental variable techniques offer great promise, but, to date, to the best of our knowledge, there have been only two applications of CACE in GBG trials.[47,48] Leveraging data from a recent US-based RCT,[37] Bradshaw et al.[47] found that the presence and magnitude of intervention effects for at-risk children in 'PATHS to PAX' (an integration of the PAX GBG and the Promoting Alternative Thinking Strategies curriculum) varied as a function of compliance. Thus, the initial effect on social competence grew from 0.01 to 0.28, and previously unidentified effects on academic engagement and emotion regulation emerged in CACE models that took account of variability in GBG dosage (e.g. total duration of exposure in minutes). However, in models focusing on the PAX GBG alone, there was no difference between (null) initial and CACE findings; in other words, the PAX GBG was found to be ineffective for at-risk children even after robustly accounting for implementation variability. The second application of CACE was reported by the authors of the current study using data from the English GBG trial[1] that is the focus of this report (see *The Good Behaviour Game in England*). In contrast to null ITT findings, Ashworth et al.[48] revealed sleeper effects of the intervention on academic attainment at the 12-month post-intervention follow-up when compliance (i.e. dosage, as in the Bradshaw study[47]) was taken into account.

Last, the effects of the GBG may be moderated by the timing of follow-up. It is important to study intervention effects over time to establish whether effects detected immediately post intervention are sustained (i.e. maintenance effects) or effects only become apparent in the years that follow (i.e. sleeper effects). Although medium-term follow-up is generally lacking, there is promising evidence of longer-term maintenance effects in the GBG. For example, when Ialongo et al.[28] followed up the sample of one of the original US trials[17] ≈ 5 years after the intervention was concluded, those who had received the GBG (in combination with curriculum enhancements and back-up strategies) were significantly less likely than those in the control condition to meet the diagnostic criteria for conduct disorder. Similarly, following up another US trial sample,[37] Kellam et al. reported that male participants who were initially classified as aggressive and had participated in the GBG in first grade (i.e. aged 6–7 years) were significantly less likely to engage in high-risk sexual behaviours and drug abuse as young adults (i.e. aged 19–21 years).[49] Such findings are in line with the LCSFT that underpins the GBG, as they are demonstrative of effective socialisation of behaviour influencing social adaptational status in other social fields as these change throughout the life course.[15]

The Good Behaviour Game in England

Two early studies of the GBG in England were published in the 1980s, but were very small scale, lacked a comparison group and focused solely on the utility of the game in increasing on-task behaviour among children and young people in special education settings.[50,51] More recently, Oxford Brookes University (Oxford, UK) led a pilot of the AIR version of the GBG in Oxfordshire over the course of a single school year in 10 classrooms (n = 222 children aged 5–9 years).[52] Although this study also lacked a control group, it established the acceptability and feasibility of the GBG in the English school context and provided tentative evidence of its impact on behavioural and other outcomes.

Subsequently, Mentor UK (London, UK) successfully applied for funding from the EEF to implement the GBG on a much larger scale.[1] The authors of the current report were appointed as independent evaluators in a major RCT[1] involving 77 primary schools in 23 local authorities across three regions of England, the findings of which are noted in *Table 1*. The EEF trial assessed the immediate impact of the GBG on reading attainment and behavioural (i.e. disruptive behaviour, concentration problems and prosocial behaviour) outcomes.[1] As noted in *Table 1*, null findings were reported. This trial also included a parallel mixed-methods implementation and process evaluation (IPE), comprising surveys, structured observations and qualitative school case studies (developed from interviews with GBG leads, teachers, headteachers, teaching assistants and parents; pupil focus groups; informal observations; and field notes). This was designed to (1) establish a clear counterfactual and give an indication of the level of programme differentiation between the GBG and usual practice, (2) document the implementation of the GBG and (3) develop a rich, detailed picture of the implementation process and the factors underpinning it.

To avoid duplication, the EEF IPE findings are not reported in detail here; instead, we recommend that the interested reader access the freely available report.[1] In summary, the IPE found that usual practice in behaviour management in participating schools included practices that mirrored some of the core components of the GBG (e.g. classroom rules, team membership, monitoring behaviour and positive reinforcement), indicating low programme differentiation. Although most aspects of implementation (e.g. fidelity, quality, reach) achieved good levels, it was notable that dosage was markedly lower than that recommended by the developer.[53] Intervention characteristics (e.g. lack of direct interaction with children during game sessions), the implementation support system (e.g. coaching support), classroom-level factors (e.g. pupil needs and attitudes, teacher attitudes) and school-level factors (e.g. school climate and openness to change) were reported to influence implementation of the GBG in participating schools. We return to some of these findings in *Chapter 4* in view of their potential to help explain the findings reported here.

As the EEF trial focused primarily on academic attainment and behavioural outcomes in the period immediately following the end of the intervention, we sought funding from NIHR to (a) augment outcome assessment to include health-related outcomes, beginning at the immediate post-intervention follow-up; (b) assess sleeper and/or maintenance effects at 12- and 24-month post-intervention follow-ups; and (c) perform an economic evaluation.

In accordance with the above, our hypotheses were as follows:

1. What is the impact of the GBG on health-related outcomes for children?

   ○ Hypothesis 1 – children in primary schools implementing the GBG over a 2-year period will demonstrate significantly better outcomes in mental health; conduct problems (hypothesis 1a), psychological well-being (hypothesis 1b) and emotional symptoms (hypothesis 1c); sources of resilience; peer and social support (hypothesis 1d) and school environment (hypothesis 1e); school absence (hypothesis 1f), and significantly lower rates of bullying (i.e. social acceptance; hypothesis 1g) and exclusion from school (hypothesis 1h), than children attending control schools.
   ○ The primary trial outcome was conduct problems (hypothesis 1a). The secondary mental health outcomes were psychological well-being and emotional symptoms; this is consistent with both the GBG logic model[52] and LCSFT.[15] Sources of resilience (e.g. peer and social support, and school environment) were included to assess the extent to which intervention exposure increases children's ability to draw on these (consistent with LCSFT).[15] Bullying (i.e. social acceptance) is included as a proxy for improved social adaptational status and positive interactions among peers (as predicted by the GBG logic model).[52] Last, school absence and exclusion from school are included to assess the extent to which improvements in the aforementioned domains translate into measurable change in school outcomes relating to engagement and behaviour (as predicted by the GBG logic model).[52]

2. Are there differential effects of the GBG for boys at risk of developing conduct disorders?

○ Hypothesis 2 – boys at risk of developing conduct disorders [defined as scoring in the borderline or abnormal ranges of the conduct problems subscale of the teacher-rated Strengths and Difficulties Questionnaire (SDQ) at baseline] in primary schools implementing the GBG over a 2-year period will demonstrate significantly better outcomes in mental health; conduct problems (hypothesis 2a), psychological well-being (hypothesis 2b) and emotional symptoms (hypothesis 2c); sources of resilience; peer and social support (hypothesis 2d) and school environment (hypothesis 2e); school absence (hypothesis 2f), and significantly lower rates of bullying (i.e. social acceptance; hypothesis 2g) and exclusion from school (hypothesis 2h) than at-risk boys attending control schools.

○ We expect amplified effects of the GBG for boys at risk of developing conduct disorders on the basis of previous research findings.[23,24] As noted above, the intervention procedures are likely to particularly appeal to boys given the gendered socialisation of competitiveness;[44] furthermore, the sex ratio for the prevalence of conduct disorders in childhood is approximately 3 : 1 in favour of boys.[4]

3. Do the effects of the GBG vary by intervention compliance?

○ Hypothesis 3 – the magnitude of intervention effects noted in hypothesis 1a–h above will vary as a function of intervention compliance. Specifically, we predict larger ESs in schools defined as compliers in terms of dosage (hypothesis 3a–h).

○ Research across multiple disciplines has consistently demonstrated that interventions are rarely implemented as designed and, crucially, that variability in implementation is associated with variability in the achievement of expected outcomes.[54] CACE and related approaches allow researchers to robustly determine treatment effects in the context of receipt of an intervention (as opposed to the offer of said intervention, as in ITT estimation, which we use in hypothesis 1 above). Initial applications of CACE in GBG trials have produced mixed results,[47,48] warranting further exploration here.

4. Are any effects of the GBG on health- and education-related outcomes sustained over time?

○ Hypothesis 4 – the effects of the GBG on mental health; conduct problems (hypothesis 4a), psychological well-being (hypothesis 4b) and emotional symptoms (hypothesis 4c); sources of resilience; peer and social support (hypothesis 4d) and school environment (hypothesis 4e); school absence (hypothesis 4f), bullying (i.e. social acceptance; hypothesis 4g) and exclusion from school (hypothesis 4h), reading attainment (hypothesis 4i), prosocial behaviour (hypothesis 4j), concentration problems (hypothesis 4k) and disruptive behaviour (hypothesis 4l) will be maintained at 12- and 24-month post-intervention follow-ups.

○ This hypothesis is based on existing evidence of the sustained effects of the GBG[28] and LCSFT that suggests that effective socialisation of behaviour can yield a lasting influence on children's social adaptational status.[15] The inclusion of an interim (e.g. 12-month) follow-up is critical to model the maintenance (or emergence) of effects with greater precision.

○ The reader will note that there are several additional outcomes for hypothesis 4 (i.e. reading attainment, prosocial behaviour, concentration problems and disruptive behaviour) compared with hypotheses 1 and 2. These outcomes are not addressed in hypotheses 1 and 2 because the analyses pertaining to them were included in the main report for the EEF-funded trial (see *Table 1*).[1]

5. To what extent are children's educational and health-related outcomes related over time?

○ Hypothesis 5 – children's educational and health-related outcomes will be related over time.

○ Drawing on developmental cascades theory,[55] we focus on cross-domain associations over time in three key areas of functioning (i.e. emotional symptoms, conduct problems and academic

attainment), as well as accounting for the potential confounding influences of shared risk [e.g. poverty, special educational needs and disabilities (SEND)], trial group and sex.[5,56] Accordingly, this aspect of our project is intended to generate insights relating to the connections between learning and mental health during the transition from middle childhood to early adolescence, as opposed to the implementation and impact of the GBG per se. We extend existing research in this area by estimating both within- and between-individual effects.

6. Can the GBG be regarded as providing value for money?

- Hypothesis 6 – the GBG will represent an efficient use of resources when considered from a public-sector perspective.
- There is good reason to propose that the GBG could prove to be an efficient use of resource,[57] but quantification of the benefit of educational interventions is challenging. The planned economic analysis allowed for assessment of both monetised and non-monetised outcomes, the hypothesis being that, on balance, the benefits of implementing the GBG will balance the costs of implementation.

Chapter 2 Methods

Parts of this chapter are reproduced or adapted with permission from Humphrey *et al.*[1] Contains information licensed under the Open Government Licence v3.0. URL: www.nationalarchives.gov.uk/doc/open-government-licence/version/3/.

Parts of this chapter are also reproduced or adapted with permission from the GBG trial protocol (available from the NIHR project web page: www.journalslibrary.nihr.ac.uk/programmes/phr/145238).

Design

A two-group, parallel, cluster-randomised controlled trial[58] design was used, with schools as the unit of randomisation. *Figure 1* depicts the flow of schools and children through the trial in line with the Consolidated Standards of Reporting Trials (CONSORT). Schools assigned to the intervention arm of the trial delivered the GBG throughout the school years 2015/16 and 2016/17. Those assigned to the control arm of the trial continued with their usual practice during this period. Data were collected at baseline [pre randomisation, time 1 (T1); May–July 2015] and on an annual basis thereafter for 4 years [i.e. time 2 (T2), time 3 (T3), time 4 (T4) and time 5 (T5)].

Ethics, approval, consent and trial monitoring

The University of Manchester Research Ethics Committee at the University of Manchester (Manchester, UK) approved the study (reference number 15126). The consent/assent process involved three stages.

First, eligible schools signed a memorandum of agreement (MoA) indicating their willingness to participate. The MoA documented the nature of participation (e.g. data collection procedures and requirements, plus payment of a contributory fee by those schools allocated to the intervention arm), the RCT design (e.g. that half of participating schools would be randomly allocated to implement the GBG) and what schools would receive 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-practice arm).

Second, participating schools sent information and opt-out consent sheets to parents of all eligible children. Parents/carers wishing to opt their children out of the study were able to do so by returning the opt-out form on the consent sheet to the research team via a Freepost address at the University of Manchester. In total, 68 parents (2.2%) exercised their right to opt their children out of the study. Parental consent was for participation in our research, as opposed to participation in the GBG itself, as the latter was determined by each school's in loco parentis responsibilities. Hence, children who were opted out in the intervention arm still took part in the GBG but did not provide data for this report.

Third, 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.

A clear monitoring and reporting function [i.e. schools to the research team, and the research team to the Trial Steering Committee (TSC)] was established. No adverse events were reported.
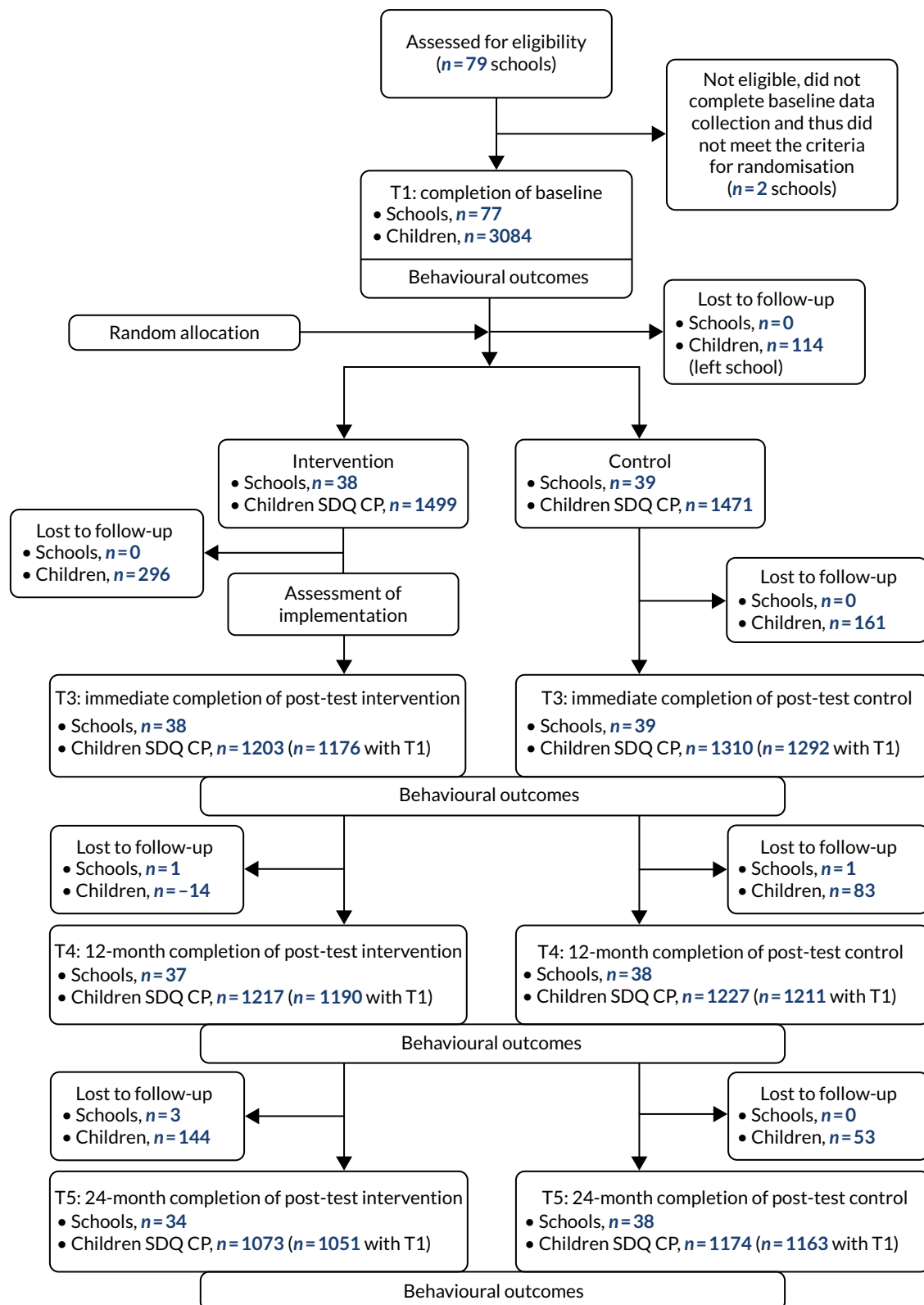
FIGURE 1 The CONSORT flow diagram of schools and children through the GBG trial for the primary outcome (conduct problems subscale of teacher informant-report SDQ CP). Note that T2 was omitted as this data point was not used in the current study. CP, conduct problems; T1, time 1; T2, time 2; T3, time 3; T4, time 4; T5, time 5.

Participants, recruitment and randomisation

Sample size and power

A total of 79 schools were recruited, of which 77 met the eligibility criteria for randomisation (i.e. signing the MoA and completing baseline measures). There were 3084 eligible children (i.e. those in Year 2 at T1) attending these schools. As this sample was initially recruited for the EEF-funded education-related trial,[1] the numbers of schools and children required were based on prospective power calculations relating to the primary outcome for that project (i.e. children's reading scores at T3). Accordingly, the power calculations presented here for the health-related analyses are necessarily post hoc.

A Monte Carlo simulation (see *Report Supplementary Material 1*), with robust maximum likelihood and 10,000 replications, was conducted in M*plus*, version 8.2 (Muthén & Muthén, Los Angeles, CA, USA), to assess the power for the GBG intervention effect on the main outcome (i.e. conduct problems), with 3084 children across 77 clusters (mean cluster size, $n = 40$). Attrition was set to 0% and 15% for the intervention and outcome variables, respectively. The binary intervention variable (GBG vs. control) was set to have a variance of 1 and a mean of 0. The simulation was carried out for a logistic regression, with different seed values to ensure the stability of findings. For an intracluster correlation coefficient value of 0.06, the between-level variance was set to 0.019 using the following formula:

$$S_b^2 = \rho / (S_w^2 - \rho), \tag{1}$$

where S_w^2 is set to $\pi^2/3$ for logistic regression. The probability of being at risk was set to 13.4% based on national averages.[59] Therefore, the threshold for the outcome variable and the slope of the trial were adjusted to reflect this using the following formula:

$$p\,(y = 1\,|\,x = 1) = 1\,/\,[1 + \exp(\log)], \tag{2}$$

where $\log = a + b \times x$. Acceptable power levels were achieved when the trial effect was set to $b = -0.38$ and the outcome threshold to $\tau = 2.25$.[60] Findings showed small population parameter (0.16%) and standard error (SE) bias (1.4%), and satisfactory coverage (0.94) and power (0.80) for an ES of $b = -0.38$, which corresponds to $\Delta = 0.38$, based on a total variance of 1 and $\Delta = b$/standard deviation (SD). This means that the trial was powered to detect intervention effects considered to be small to moderate when judged by conventional thresholds,[42] aligning well with the overall trend reported across previous GBG trials (see *Chapter 1, Evidence base for the Good Behaviour Game*).

Recruitment of schools

Mentor UK recruited schools to the trial from three regions (i.e. Greater Manchester, West and South Yorkshire, and the East Midlands) using a number of strategies, including regional recruitment events, using contacts at local authorities and independent providers to identify prospective trial schools, and e-mailing project flyers to schools. Initial expressions of interest were sought using an online form, followed by direct contact from Mentor UK, before the MoA was signed. Recruited schools were then assessed for representativeness against all schools on key characteristics (e.g. size, FSMs) (see *Table 3* and accompanying commentary).

Inclusion/exclusion criteria

All children in Year 2 (i.e. aged 6–7 years) at T1 (May–July 2015) were eligible to participate.

Randomisation

Participating schools were the unit of randomisation to minimise the risk of contamination that would have resulted from within-school (e.g. class) randomisation, and for practical reasons, given that the intervention model includes a GBG coach being assigned to each participating school in the intervention arm.

A total of 77 schools met the criteria for randomisation (i.e. signed MoA, > 90% of T1 measures complete) and were randomly allocated to either implement the GBG or continue usual practice following the completion of baseline measures at T1. The allocation procedure was conducted independently by the Manchester Academic Health Science Centre Clinical Trials Unit. A minimisation algorithm was applied to the randomisation to ensure balance across the arms of the trial in terms of the proportion of children eligible for free school meals (FSMs) and school size (data were provided from the school performance tables on the Department for Education website.[61] This approach is described as the 'platinum standard' for trials, conferring the benefits of randomisation in terms of rigour and causal inference, as well as guaranteeing similarity of groups on key observables.[59] Thirty-eight schools were allocated to implement GBG and 39 schools were allocated to continue with their usual practice.

Intervention

For clarity and transparency, we describe the GBG in detail using an adapted version of the Template for Intervention Description and Replication (TIDieR)[62] for use with school-based interventions,[63] thereby mirroring reporting in the education-related trial that set the foundation for the current study.[1]

Brief name
The Good Behaviour Game (GBG).

Why (rationale/theory)
The GBG is underpinned by three theories of human development and learning: behaviourism (specifically contingency management),[13] social learning theory[14] and LCSFT.[15] In terms of behaviourism, it is assumed that behaviour that is rewarded is more likely to be reproduced. Consequently, children receive positive reinforcement when they engage in desired behaviours (e.g. following the teacher's instructions during an activity). However, the group-based orientation of the GBG means that it also draws on social learning theory. In particular, at-risk children can benefit from appropriate behaviour being modelled effectively by other team members. Last, a key tenet of LCSFT is that successful adaptation at different stages of life is contingent on our ability to meet particular social task demands. In school, these 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 a given social field (e.g. teachers, peers). LCSFT predicts that improving the way in which children are socialised in the classroom (e.g. explicitly highlighting and promoting social task demands, then rewarding children for meeting them) will improve their social adaptation. It is also predicted that early improvements in social adaptation in the classroom will extend to positive adaptation in other social fields (e.g. peer group, family, work) throughout the lifespan.[15]

Who (recipients)
The GBG is a universal intervention that is delivered to all children in a given class.

What (materials)
Schools receive GBG manuals that outline 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 in to a secure website to record game and probe data [see *What (procedures)*] in real time and retrospectively, and these data could then be downloaded to assess temporal trends and inform future implementation planning. In turn, each GBG coach [see *How well (planned)*] was able to access their assigned teachers' data for use in later support sessions, and the research team was able to access all teachers' data so that it could be used to monitor the length and frequency of games (used in this study to examine the extent to which intervention effects varied by levels of compliance; see *Hypothesis 3: implementation effects*).

Second, we developed an electronic version of the fidelity checklist used by GBG coaches. This was identical to the paper version used by the licensing organisation (AIR) and was used for the same purpose (e.g. to facilitate feedback following an observation session).

What (procedures)

The teacher divides the class into mixed teams with up to seven members, with team membership typically varied several times in a school year (e.g. every half term). Where possible, each team should be balanced, with equal representation of salient factors such as behaviour, academic ability and sex. 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 of the following four rules among the teams:

1. We will work quietly.
2. We will be polite to others.
3. We will only get out of our seats with permission.
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 sitting next to you can hear you), level 2 (inside voice, only people sitting 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, which then access an agreed reward.[15,52] The procedures undertaken before, during and immediately after a game session are detailed in the aforementioned intervention manual [see *What (materials)*] and are as follows:

- Before the game –

  - The teacher explains the task/activity.
  - The teacher checks understanding of the task/activity.
  - The teacher reminds pupils that they cannot ask for help.
  - Pupils are placed in teams of between 3 and 7 (except in special circumstances, 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).
  - The pupils are in clear teams.
  - The teams are sex balanced.
  - The 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.
  - The voice level for the task/activity is given by the teacher.
  - The teacher states when the game begins.
  - The teacher states how long the game will be played for.
  - The teacher sets a timer.
  - The teacher states that they will monitor infractions.
  - The teacher states that four infractions are permitted per team.
  - The teacher reminds pupils that they are not competing against each other.

- During the game –

  - The teacher identifies infractions when they occur.
  - The teacher records infractions on the scoreboard.
  - The teacher identifies rule breaking team (e.g. 'team 4 have broken rule 4: "we will follow directions" ').

- ○ The teacher discreetly indicates the infraction to specific pupil.
- ○ The rest of the team and/or class are praised for adhering to rules (e.g. 'well done everyone else for following rule 4').
- ○ The teacher does not punish pupils/teams for infractions.
- ○ The teacher monitors behaviour.
- ○ The teacher does not interact with pupils.
- ○ The teacher adheres to the time limit that was set.
- ○ The teacher announces the end of the game.

- After the game –

  - ○ The teacher repeats the criterion of four infractions or fewer.
  - ○ The teacher announces the winning team(s) only.
  - ○ Members of the winning team receive a stamp (or marker, etc.) in their individual booklets.
  - ○ A star is placed on the wall chart (or equivalent).

Over the course of the implementation of the GBG, it is intended for there to be a natural progression in terms of the types of rewards used (from tangible rewards, e.g. stickers, to more abstract rewards, e.g. free time), how long the game is played for (from 10 minutes to a whole lesson), the frequency at which the game is played (from three times per week to every day) and when rewards are given (at the end of the game, end of the day or at the end of the week).[11,64] This progression is designed to maintain responsiveness, interest and challenge for students, as well as 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 (i.e. maintenance) and will be exhibited in all settings (i.e. generalisation). These processes are documented through 'game' and 'probe' data collected by teachers during implementation.[52] Probe data, used to assess generalisation, are collected covertly during an ordinary task/activity, following the same procedures as those used in a game session (e.g. the teacher monitoring rule infractions among teams), but without explicitly setting up the rules and announcing infractions.

Who (provider)
The GBG is implemented by class teachers.

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, the GBG 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 *What (procedures)*] mean that some time is taken up before and after the game period/class activity.

Where
The GBG is implemented on site in participating schools.

When and how much
The GBG is played throughout the school year. As in *What (procedures)*, 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 per week to every day).

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 the fidelity of implementation.

However, it is now widely accepted that some form of adaptation is inevitable and may be desirable to improve local ownership and fit to context.[65,66] A critical aspect of the GBG coach role, therefore, is to support teachers to make adaptations that are in keeping with the goals and theory of the intervention.[67]

How well (planned)
Teachers receive 3 days of training (2 days of initial training and 1 day of follow-up training approximately 4 months later) from coaches (mostly former teachers), who are contracted by Mentor UK and trained by AIR. Day 1 of the initial training covers an introduction to the GBG theory and logic, such as understanding the core elements of the game (e.g. class rules, team membership, positive reinforcement and monitoring). Day 2 focuses on implementation procedures and practices (e.g. overview of successful GBG implementation, introduction to the implementation fidelity checklist and development of a plan for implementation for their class). The follow-up 'booster' training session revisits these ideas, with an opportunity for sharing of good practice and problem-solving.

Ongoing technical support and assistance is provided by the trained coaches, as noted in *Tailoring*. In the current study, participating schools were each allocated a GBG coach who visited approximately once per month to support implementation throughout the trial. These visits typically comprised modelling of game sessions, observation and feedback [including review of the game, and the probe and fidelity checklist data – see *What (procedures)*], ad hoc e-mail and telephone support, and provision of additional/booster training or information sessions, as required.

Usual school practice

Schools allocated to the control arm of the trial continued their usual school practice during the main trial period (T1–T3). To better understand the nature of this practice, and thereby establish a robust counterfactual, a survey of teachers' behaviour management strategies and approaches[68] was administered to teachers at T1 and T2. The findings of this survey are reported in detail in the main report for the education-related trial[1] and so we present headline findings only. The following were taken from the T1 control group data:

- 'I establish and maintain a set of classroom rules': 95.1% endorsed 'yes'.
- 'I communicate clear expectations about rules and pupils' responsibilities, e.g. through posters': 90.2% endorsed 'yes'.
- 'I observe and monitor pupils' behaviour in the classroom': 100% endorsed 'yes'.
- 'I use prizes as rewards for good behaviour': 59.9% endorsed 'weekly' or 'every day'.
- 'I use group rewards': 66.6% endorsed 'weekly' or 'every day'.

These data appear to indicate relatively low programme differentiation; in other words, teachers in the control arm of the trial were enacting behaviour management practices that mirrored some of the core components of the GBG (e.g. classroom rules, team membership, monitoring behaviour and positive reinforcement). However, as a counterpoint, we note that the idea of an 'untreated' control group in the context of school-based preventative interventions has long been regarded as a fantasy,[45] and certain practices that are core to the GBG (e.g. establishing and maintaining a set of classroom rules) are so endemic that one would be hard pressed to find a classroom setting in which they are completely absent.

Assessment of outcomes

In selecting our primary and secondary outcomes measures, we used the following criteria: (1) goodness of fit with study parameters (e.g. age of participants, domains of interest); (2) psychometric properties (using the thresholds set by Terwee[69]); (3) brevity and accessibility; and (4) use in similar or related

research published in peer-reviewed journals (e.g. the measure has been used in a previous RCT of a school-based preventative intervention). The measures were approved by the TSC, including our patient and public involvement (PPI) experts, Common Room [further information about Common Room can be found at URL: https://commonroom.uk.com/ (last accessed November 2021)]. *Table 2* provides a summary of the measures and informants/data sources at T1 and T3–T5 (note that no data collected at T2 are used in this study).

Primary outcome measure

Conduct problems
The primary outcome measure for the trial was the conduct problems subscale of the teacher informant-report version of the SDQ,[60] for which we also had pre-test data, as this was used to identify the at-risk sample at baseline. It comprises five items, for which respondents read a statement (e.g. 'often has temper tantrums or hot tempers') and indicate their agreement on a three-point scale (i.e. not true, somewhat true and certainly true). The scale has a possible range of 0–10. Previously, the teacher informant-report SDQ has been shown to exhibit acceptable reliability (internal Cronbach's alpha of up to 0.87; test–retest

TABLE 2 Outcomes assessed in the GBG trial

Time point	Child self-report or test	Teacher informant report	National Pupil Database
Baseline (T1)		Conduct problems (SDQ)	School absence
		Disruptive behaviour, concentration problems and prosocial behaviour (TOCA-C)	Exclusion from school
			KS1 reading attainment
			Child-level covariates (e.g. sex, FSMs eligibility)
End of main trial period (T3)	Psychological well-being (Kidscreen-27)	Conduct problems and emotional symptoms (SDQ)	School absence
	Bullying (i.e. social acceptance) (Kidscreen-52)		Exclusion from school
	Peer and social support, and school environment (Kidscreen-27)		
12-month post-intervention follow-up (T4)	Hodder Group Reading Test	Conduct problems and emotional symptoms (SDQ)	School absence
	Psychological well-being (Kidscreen-27)	Disruptive behaviour, concentration problems and prosocial behaviour (TOCA-C)	Exclusion from school
	Bullying (i.e. social acceptance) (Kidscreen-52)		
	Peer and social support, and school environment (Kidscreen-27)		
24-month post-intervention follow-up (T5)	Psychological well-being (Kidscreen-27)	Conduct problems and emotional symptoms (SDQ)	School absence
	Bullying (i.e. social acceptance) (Kidscreen-52)	Disruptive behaviour, concentration problems and prosocial behaviour (TOCA-C)	Exclusion from school
	Peer and social support, and school environment (Kidscreen-27)		KS2 reading attainment

KS, key stage; TOCA-C, Teacher Observation of Child Adaptation Checklist.

r of up to 0.8) and validity [factorial, established through confirmatory factor analysis (CFA); convergent, correlates with a range of similar instruments; predictive, strongly predictive of independently diagnosed psychiatric disorders).[60] Reliability of the conduct problems scale in the current study at T1 was Cronbach's alpha of 0.80.

Baseline scores on this measure were used to identify our at-risk sample for hypothesis 2. A score of 0–2 represents the normal range, 3 represents borderline and 4–10 represents the abnormal range.[60] At-risk status was defined as scoring in the borderline or abnormal range on this measure at T1.

Secondary outcome measures

Psychological well-being
The Kidscreen-27 psychological well-being subscale provides a self-reported assessment of children's mental health.[70] It is brief, comprising seven items in which respondents read a statement (e.g. 'thinking about last week, have you been in a good mood?') and indicate their agreement on a five-point scale (i.e. never, seldom, quite often, very often and always). The Kidscreen-27 was designed and validated for use with children aged $\geq$ 8 years. Previously, the measure has been shown to exhibit good internal consistency (α coefficient 0.84), a robust factor structure (established through CFA) and strong predictive validity [e.g. discriminates between those identified with mental health problems, as assessed by the SDQ (ES 0.68), and correlates with similar measures, e.g. the Youth Quality of Life Instrument (ES 0.63), Child Health questionnaire (ES 0.36) and Child Health and Illness Profile (ES 0.62)[70]].

Emotional symptoms
The teacher informant-report version of the SDQ emotional symptoms subscale comprises five items in which respondents read a statement (e.g. 'many worries, often seems worried') and indicate their agreement on a three-point scale (i.e. not true, somewhat true and certainly true); the scale has a possible range of 0–10.[60] As noted above (see *Conduct problems*), the teacher informant-report SDQ has previously been shown to exhibit acceptable reliability and validity.

Bullying (i.e. social acceptance)
The Kidscreen-52 social acceptance domain provides a self-reported assessment of experience of bullying among children.[70] It is brief, comprising three items in which respondents read a statement (e.g. 'thinking about last week, have other girls and boys made fun of you?') and indicate their agreement on a five-point scale (i.e. never, seldom, quite often, very often and always). The Kidscreen-52 was designed and validated for use with children aged $\geq$ 8 years. Previously, it has been shown to exhibit good internal consistency (α coefficient 0.77), a robust factor structure (established through CFA) and strong predictive validity (e.g. discriminates between those identified with mental health problems, as assessed by the SDQ).[70]

Sources of resilience
The Kidscreen-27 social support and peers and school environment domains provide a self-reported assessment of sources of resilience among children.[70] They are brief, each comprising four items in which respondents read a statement (e.g. social support and peers, 'thinking about last week, have you spent time with your friends?'; school environment, 'thinking about last week, have you been happy at school?') and indicate their agreement on a five-point scale (i.e. social support and peers: never, seldom, quite often, very often and always; school environment: not at all, slightly, moderately, very and extremely). The Kidscreen-27 was designed and validated for use with children aged $\geq$ 8 years. As noted in *Psychological well-being*, the measure has previously been shown to exhibit acceptable reliability and validity.[70]

Reading attainment
The baseline period (T1) for the trial coincided with the end of key stage (KS)1 teacher assessments for the trial cohort, so children's KS1 National Curriculum reading point score (i.e. the KS1_READPOINTS variable) was used as a pre-test covariate. This was extracted from the National Pupil Database (NPD)

at baseline.[71] Assessment of reading attainment at the 12-month follow-up (T4) used the Hodder Group Reading Test (HGRT), test sheet 2, which is suitable for pupils aged 7–12 years.[72] This paper-based measure produces raw scores, which are used in the analyses reported here, but can also be transformed into National Curriculum levels, reading ages and standardised scores. The HGRT is administered in a whole-class/group context and takes ≤ 30 minutes to complete. Assessment of reading attainment at the 24-month follow-up (T5) coincided with the end of KS2 statutory assessment tests for the trial cohort, so children's KS2 National Curriculum Reading scaled score (i.e. the KS2_READSCORE variable) was drawn from the NPD.

Behaviour

Children's behaviour was assessed using the Teacher Observation of Children's Adaptation checklist (TOCA-C).[73] This 21-item scale provides indices of children's concentration problems, disruptive behaviour and prosocial behaviour. Teachers read statements about a child (e.g. 'pays attention') and endorse them on a six-point scale (i.e. never, rarely, sometimes, often, very often and 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. The prosocial 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 sex, race and age.[73]

School absence and exclusion from school

Data on children's school absence and exclusion from school were extracted from the NPD at baseline (T1), the immediate follow-up (T3), and at the 12-month (T4) and 24-month (T5) follow-ups. For absence data, to allow for maximum variation and a continuous measurement scale, the number of sessions children were absent for during the academic year (variable OverallAbsence_6HalfTerms_ab[yy]) and the number of sessions possible for the academic year (variable SessionsPossible_6HalfTerms_ab[yy]) were extracted, allowing the proportion of overall absence to be calculated. The overall number of sessions absent, a combination of authorised (i.e. a valid and acceptable reason has been provided and approved by the school) and unauthorised (i.e. any absence that the school has not given permission for or where an explanation has not been provided) absences was used as the outcome variable, as we were interested in the effects of the intervention on any time spent away from school, irrespective of the reason for this. In addition, schools are known to vary in their interpretation of authorised absence (e.g. holiday leave during term time). Exclusion data were extracted, detailing the total number of sessions for fixed exclusions for the academic year (variable TotalFixedSessions_ex[yy]).

Covariates

Background data on both schools (e.g. school size, proportion of children eligible for FSMs) and children (e.g. sex, FSM eligibility) were collected for use as covariates in our analyses. School-level data were taken from Department for Education performance tables, and child-level data were extracted from the NPD. The NPD also provides an anonymised child reference number that was used to ensure accurate data matching (e.g. across time and between informants).

Assessment of implementation

Assessment of implementation was undertaken to determine the extent to which intervention outcomes varied as a function of compliance in dosage (i.e. how frequently is the GBG played and for how long?). Dosage data were generated using a bespoke, online 'scoreboard' tool designed by the research team.[1] This online scoreboard was introduced after the October half-term in the first year of the trial. Therefore, in the first year of the project, the scoreboard was not fully embedded until January. In total, 31 of the 38 GBG schools used the online scoreboard at least once during the year, across 49 of the 60 classes. By the second year of the trial, all implementing schools and classes were using the online scoreboard from the outset. Each teacher was able to log in to a secure website to record game and probe data in

real time and retrospectively, which could then be downloaded to assess temporal trends and inform future implementation planning. In turn, each GBG coach was able to access their assigned teachers' data for use in later support sessions, and the research team was able to access all teachers' data so that they could be used for dosage analyses in the IPE. Therefore, how often the game was played and the duration of each session were recorded, and the total minutes played could be calculated per class. This cumulative intervention intensity,[74] defined as the total number of minutes' exposure to the intervention from T1 to T3, was used to indicate dosage.

Statistical analysis

A statistical analysis plan (SAP) was developed by the research team, with support from members of the TSC with statistical expertise, and signed off by the TSC chairperson (Tamsin Ford), prior to any analyses being undertaken (8 February 2019). The full SAP can be accessed on the NIHR project page[75] and includes syntax for the analyses reported herein; in this section, we provide a basic overview of our analytical procedures. We also note deviations from the SAP, outlining their nature and rationale. Syntax for models not included in the SAP (e.g. those fitted following a deviation from the SAP) can be found in *Report Supplementary Material 1*. Analyses were undertaken using M*plus*, version 8.4 (Muthén & Muthén), Stata, version 16.1/2 (StataCorp, College Station, TX, USA), the R package 'lavaan', version 0.6.4 (The R Foundation for Statistical Computing, Vienna, Austria)[76] and M*plus*Automation (Muthén & Muthén).[77] The analysis and presentation of data follow CONSORT guidelines in relation to RCTs where applicable.[78]

Procedures for handling missing data

First, the proportion of missing data was determined for a given outcome variable. If, for a given analysis, < 5% of data were missing then a complete-case analysis was undertaken. Second, if > 5% of data were missing, differences between partially and completely observed cases were examined to establish any pattern to the missingness. Multilevel logistic regression was used to predict missingness, whereby each child was coded as having non-missing (0) or missing (1) outcome data, with other study data as explanatory variables (e.g. trial group, KS1_READPOINTS, conduct problems and TOCA-C behaviour scores at T1, sex, and FSMs). Third, if this analysis determined that data were likely to be missing at random (i.e. conditional on other observed variables), then full information maximum likelihood (FIML) estimation was used so that partially and completely observed cases of all 77 schools and 3084 children were included in our analyses, thereby reducing the bias associated with attrition.

Hypothesis 1: intention-to-treat effects

Summary

We fitted statistical models to determine the main effects of the GBG on our trial outcomes. The different models we used took account of the type of outcome data (e.g. binary, continuous), the fact that the data were nested (e.g. children within schools) and sometimes skewed (e.g. very few children are ever excluded from school), and a range of potential confounds (e.g. sex) that might also be related to our outcomes. Each model allowed us to estimate the size of any intervention effect on a given outcome, and whether or not this finding was statistically significant.

For the ITT analysis of the primary trial outcome (i.e. conduct problems), a two-level (i.e. school, child), random-intercepts logistic regression model was fitted, with post-trial (T3) conduct problems status (for which 0 = normal, 1 = borderline/abnormal) as the response variable. Trial group (i.e. GBG vs. usual practice) and minimisation variables (i.e. per cent FSMs, size) were fitted at the school level; T1 conduct problems score, sex and FSM eligibility were fitted at the child level. An intervention effect was noted if the coefficient associated with the school-level trial group variable was statistically significant. In the case of conduct problems, a significant negative coefficient would indicate reduced odds of scoring in the borderline/abnormal SDQ band at T3 as a result of allocation to the GBG arm.

For ITT analyses pertaining to secondary outcomes, a post-test-only design was used, with the exception of school absence and exclusion from school, for which baseline data were available from the NPD. For emotional symptoms, the analysis mirrored that undertaken for the primary outcome. For school absence and exclusion from school, negative binomial multilevel models were fitted because of the fact that data were in count form and highly positively skewed (e.g. there were very low counts of children absent or excluded from school). Moreover, models for school absence were controlled for exposure (i.e. number of possible sessions). For the remaining (continuous) secondary outcomes, two-level hierarchical linear regression models were fitted. In all cases, school- and child-level variables were fitted as for the primary outcome analysis. For the multilevel linear regression models, the standardised ES, Hedges' g, was calculated using the coefficient of the trial group variable divided by the total SD of the model $\beta_{trial}/(\sigma_e + \sigma_y)$.[79] The 95% confidence intervals (CIs) were calculated as the ES $\pm$ the product of the critical value of the normal distribution ($\approx$ 1.96) and the SD of the theoretical distribution of the ES.[80] For ITT analyses pertaining to count data (i.e. absences and exclusions), ES estimates are derived from the incidence rate ratios (IRRs) and corresponding 95% CIs. IRRs correspond to the exponentiated values of the log-counts estimated via the multilevel negative binomial models. An IRR of < 1 for the school-level trial group variable would indicate that the incidence of the outcome of interest (i.e. absences or exclusions) in the GBG group is lower than that in the control group by the estimated rate.

Hypothesis 2: subgroup effects

Summary
We extended the statistical models outlined above to determine the specific effects of the GBG on boys at risk of developing conduct problems. This involved adding terms to the models that would allow us to isolate any such effects (e.g. for boys whose baseline score indicated at-risk status in GBG schools).

For our planned subgroup analyses for boys exhibiting borderline/abnormal levels of conduct problems at baseline, the models outlined above for hypothesis 1 were extended to include the following cross-level interaction terms: trial group*risk status; trial group*sex; risk status*sex; and the three-way interaction of interest for hypothesis 2, trial group*risk status*sex (e.g. if GBG, if at risk, if male). An intervention effect at the subgroup level was noted if the coefficient associated with this interaction term was statistically significant ($p < 0.05$).

Hypothesis 3: implementation effects

Summary
We fitted statistical models that enabled us to determine the effects of the GBG on our trial outcomes when teachers delivered enough of the intervention to be considered 'compliers'. Given the lack of a universally agreed threshold for what constitutes 'enough' delivery, we modelled two scenarios (50th and 75th percentiles in overall dosage defined as moderate and high compliance, respectively). CACE models use this information to compare outcomes between compliers in GBG schools and 'would-be' compliers in usual-practice schools (e.g. those who would probably have complied had they been allocated to deliver the intervention). As with the models described above, CACE allows us to estimate the size of any complier effect on a given outcome, and whether or not this finding was statistically significant.

Complier-average causal effect estimation was employed to determine whether or not the presence and/or magnitude of intervention effects noted in relation to hypothesis 1 changed once intervention compliance was taken into consideration. For each outcome, CACE models were calculated probabilistically as two-level mixture models using robust maximum likelihood.[81] T1 scores for conduct problems and child-level covariates (i.e. sex, FSM eligibility, SEND, KS1 reading scores, concentration problems and prosocial scores at T1) were fitted at the pupil level. The trial group variable (i.e. GBG vs. usual practice), along with school-level covariates [per cent FSMs, size, per cent with English as an additional language (EAL), and school-average conduct problems and KS1 scores at T1], was modelled at the school level.

The child- and school-level covariates were also used as predictors of the latent class compliance variable to increase power to detect CACE effects and decrease sensitivity to violation of assumptions, such as the exclusion restriction.[82]

The classroom-level dosage (i.e. total minutes played) in intervention schools was used to represent the categorical latent class variable (i.e. compliers vs. non-compliers). However, given the absence of a verified implementation cut-off, and following previous work on the GBG,[82] a sensitivity analysis was undertaken to compare the results from two cut-offs. Moderate compliance was used to represent classrooms that fell above the 50th percentile on dosage, and high compliance was used to represent classrooms that fell above the 75th percentile. Given that classroom-level information is not available for the control schools, thus preventing us from applying a three-level CACE analysis, classroom-level dosage was disaggregated to the pupil level.[83] For each model, 2000 random sets of starting values were generated and 500 optimisations were carried out in the final stage. The best log-likelihood value (i.e. the lowest value) required replication in at least two final-stage solutions, but preferably more than two.[84] TECH8 was monitored to ensure that the model converged and that the log-likelihood reached a stable maximum. If this failed, a given model was re-examined using twice the number of random starts (i.e. STARTS = 4000, 1000). This was to ensure that the best log-likelihood value was replicated and a global solution was reached.

For the estimation of CACE models, we assumed that treatment assignment was random, the potential outcomes of each child were not affected by the treatment status of others (the Stable Unit Treatment Value Assumption), there are neither always-takers nor defiers, and the treatment effect was 0 for those who did not participate (i.e. the exclusion criterion).[85]

In the case of school absence and exclusion from school, the CACE models were specified as single-level generalised mixture models for Poisson-distributed data, with clustered, robust SEs at the school level. The covariates were the same as those in the CACE models for continuous and binary outcomes, with the addition of an exposure variable for school absence (i.e. number of possible sessions), as well as baseline (T1) school absence and exclusion from school.

Hypothesis 4: longer-term effects

Summary
The models fitted in relation to hypothesis 4 were the same as those for hypothesis 1, except that we used outcome data collected at T4 (12 months post intervention) and T5 (24 months post intervention), instead of at T3 (immediately post intervention).

Analyses for hypothesis 4 mirrored those of hypothesis 1; in other words, we undertook an ITT analysis of each outcome variable, using models and ES metrics appropriate to the data type and distribution, with trial group and minimisation variables fitted at the school level, and baseline score and other covariates fitted at the child level. For each outcome, separate models were fitted for the 12- and 24-month post-intervention follow-ups.

Hypothesis 5: temporal associations between mental health and academic attainment

Summary
We fitted statistical models that allowed us to assess the associations between mental health outcomes (e.g. emotional symptoms and conduct problems) and academic attainment (e.g. reading scores) over time (e.g. the association between emotional symptoms at T3 and reading attainment at T4). These models took account of the stability of each outcome over time (e.g. levels of conduct problems at T4 are related to levels of conduct problems at T5), the relationship between outcomes at each time point (e.g. the association between emotional symptoms at T3 and conduct problems at T3), the fact that data were clustered (e.g. children within schools) and a range of potential confounds

(e.g. shared risk). Of importance is the fact that the models we used were able to separate genuine 'within-person' effects (e.g. a child experiencing higher levels of emotional symptoms at T3 is more likely to attain lower reading scores at T4) from trait-like 'between-person' effects (e.g. children who experience higher levels of emotional symptoms tend to attain lower reading scores).

We fitted a series of structural equation models (SEMs) of the temporal associations between conduct problems, emotional symptoms, and reading attainment, over three time points: T3, T4 and T5. The SEMs corresponded to a further extension of the random-intercept cross-lagged panel model (RI-CLPM),[86] in which we combined observed and latent cross-lagged variables. Conduct problems and emotional symptoms were treated as latent variables with ordinal indicators; hence, we used the weighted least squares of means and variance adjusted (WLSMV) estimator. By contrast, academic attainment was treated as an observed variable measured through standardised tests. Following Moilanen *et al.*,[56] the effects of cumulative shared risk (i.e. FSMs, SEND; coded 0, representing the presence of no risk exposure, 1, representing exposure to a single risk factor, or 2, representing exposure to both risk factors) were assessed through the comparison of nested models. Trial group (i.e. GBG vs. usual practice) was added as a covariate to account for any intervention effects, as was sex, given the noted differences in these kinds of cascades at this point in development.[5] At first, these variables were included in the model, although their effects were constrained to 0 (hypothesis 0 model). Using the chi-squared difference test, this model was compared with a less constrained model (hypothesis 1) in which the effects of shared risk, sex and trial group were freely estimated. A statistically significant chi-squared test would indicate that the constraints worsen the model fit (i.e. model hypothesis 1 has a better fit). Longitudinal measurement invariance was conducted for the latent variables prior to the examination of the SEM panel model. Clustering was taken into account through the estimation of clustered robust SEs (e.g. cluster = 'school_id'), given the current relative inflexibility of the RI-CLPM approach to being extended to higher-level structures.

Hypothesis 6: value for money

Summary

We presented a balance sheet of the costs and outcomes of the GBG study, comparing the schools and children who had the intervention with those who did not. The costs were tallied based on data provided by Mentor UK. These costs included those of school recruitment, training, programme delivery and GBG materials. The outcomes were based on the GBG study analysis and included educational, behavioural and well-being indicators. If a child was excluded from school, the cost of child care was added to the tally of costs. Health outcomes and service use were not measured.

A cost–consequences analysis (CCA) was conducted from a public-sector perspective. GBG implementation costs were obtained from Mentor UK in an aggregated data set describing the costs accrued, including training costs, travel costs, staff and school recruitment costs and intervention material costs. Monetised outcomes were limited to the consideration of exclusion from school. Exclusions were costed based on national tariffs reflecting the median wage of parents/caregivers (long-term cost implications were not included). Costs were analysed over the 2-year intervention period. Long-term costs and outcomes were not captured. The analysis was based on the ITT population. Subgroup analyses were not conducted as resource and cost data were provided in aggregate form only. The costs and consequences of GBG implementation were presented in tabular format. The total cost, average cost per school and average cost per pupil were reported. Consequences were limited to study outcomes that were reported as change from baseline to 12- or 24-month follow-up. Outcome data were taken from the main study analysis. Additional analysis of study outcomes data was not conducted as part of the economic analysis.

Deviations from the statistical analysis plan

Deviations from the SAP were minimised wherever possible. However, some changes were, inevitably, necessary, primarily because of two factors. First, planned analyses involving NPD data had to be altered to reflect the software available in the virtual laboratory environment of the Office for National

Statistics' (ONS) Secure Research Service (SRS) and/or ONS reporting restrictions.[87] Second, for hypothesis 6 specifically, cost data for the GBG provided by Mentor UK were not as granular as had been planned (i.e. they were provided in aggregate form only), limiting the analyses that could be undertaken. This was caused by staff turnover and the subsequent collapse of Mentor UK.

Specific SAP deviations and an accompanying rationale for each are delineated below.

Hypothesis 1
The multilevel negative binomial models for school absence and exclusion from school (i.e. hypothesis 1f and hypothesis 1h) were fitted in Stata, version 16.1, because M*plus* is not available in the ONS SRS.

Hypothesis 2
As for hypothesis 1, the models for school absence and exclusion from school (hypothesis 2f and hypothesis 2h) were fitted in Stata, version 16.1, because M*plus* is not available in the ONS SRS. Furthermore, hypothesis 2h could not be reported as planned because the number of exclusions from school in the follow-up period, split by trial arm, fell below the ONS reporting threshold of 10 individuals.[87] Instead, we report results for the 12-month post-intervention follow-up for exclusion from school.

Hypothesis 3
Complier-average causal effect models for school absence and exclusion from school (hypothesis 3f and hypothesis 3h) could not be implemented as planned because M*plus* is not available in the ONS SRS. As multilevel CACE is not possible in Stata, version 16.1, we instead fitted a single-level CACE for Poisson-distributed data, with robust SEs clustered at the school level. The single-level CACE model is the approach explained in Skrondal and Rabe-Hesketh[88] and implemented in Stata, version 16.1 (using the gsem command), by Troncoso.[89] Using the data from Little and Yau,[90] we compared the results of our Stata code with the results obtained in M*plus*,[91] Latent GOLD® (Statistical Innovations Inc., Arlington, MA, USA)[92] and the generalised linear latent and mixed models package (GLLAMM). In all four packages, the CACE estimate was nearly identical [e.g. differences in rounding only; Little and Yau,[90] –0.309 (the 'true' value); GLLAMM, –0.3098673; Stata,[89] –0.3098673; Latent GOLD,[92] –0.3099; M*plus*,[91] –0.310]. In the case of exclusion from school (hypothesis 3h), results could not be reported because the models did not reach convergence. This is most likely to be due to the extremely low counts/variability of exclusions in our sample.

Owing to preliminary convergence issues, likely to be caused by the excessive number of dimensions of integration (total of 7, with over 500 integration points), FIML was not possible in the CACE models fitted in M*plus* (i.e. for those models that could be fitted outside the SRS; hypotheses 3a–e and hypothesis 3g). Complete data sets using listwise deletion were therefore utilised. The models for school absence (hypothesis 3f) that were fitted in Stata used an equation-wise deletion procedure, using all of the available information per equation in the model (i.e. the compliance model and regression model).

Hypothesis 4
As in hypothesis 1 and hypothesis 2, the multilevel negative binomial models for school absence and exclusion from school (hypothesis 4f and hypothesis 4h) were fitted in Stata, version 16.1, because M*plus* is not available in the ONS SRS. The multilevel model for KS2 reading attainment scores (hypothesis 4i) at the 24-month follow-up was also fitted in Stata, version 16.1, for the same reason.

Hypothesis 5
These models were fitted using the R package lavaan because M*plus* is not available in the ONS SRS. In addition, the model specification deviates from the original plan, insofar as the RI-CLPM approach used in this report is fundamentally different from the traditional cross-lagged panel model (CLPM) that was originally planned. Crucially, the RI-CLPM estimates within- and between-individual effects,

whereas the CLPM approach estimates the former only. As noted by Hamaker *et al.*,[86] the CLPM makes the, often untenable, assumption that there is no between-subject variability of stable traits. In our case, there is evidence of time-invariant characteristics shaping academic attainment, rendering outcomes such as reading attainment relatively stable over time.[93–95] It would also be reasonable to assume stable traits in the case of mental health outcomes, such as conduct problems and emotional symptoms.

Furthermore, we initially set out to test a third model that incorporates sources of resilience (e.g. peer and social support) as moderators of the relationship between risk markers and outcome variables. However, this analysis route turned out to be problematic because of software limitations. As the available software solutions in the SRS (i.e. the R package lavaan) support only limited information estimators for categorical data (here, WLSMV), including time-varying covariates (e.g. peer and social support) resulted in the listwise deletion of about one-third of the sample. Given the nested nature of the modelling process, this listwise deletion was performed for all models, including hypothesis 0 and hypothesis 1, which resulted in a considerable loss of power, and, consequently, some coefficients did not reach statistical significance when they would have otherwise (e.g. in the absence of the peer and social support variables). However, this does not undermine the value of the models presented here, as we did estimate a reliable trial effect, as well as sex and shared risk effects.

Hypothesis 6

The original intention to use an individual-school budget perspective was adapted in line with other published reports to include a broader public-sector perspective.[96] Furthermore, the SAP stated that we would provide a detailed descriptive summary of implementation-related resource use, but these data were not available from Mentor UK. Relatedly, the plan to estimate variance in cost, including interquartile ranges and 95% CIs, was not undertaken as cost data were provided in aggregate form only. The SAP also stated that we would conduct a sensitivity analysis applying a monetary cost to teacher time based on GBG dosage, but this analysis was not conducted as it would double count teaching costs (i.e. no additional teacher time was required for the intervention). Last, we initially planned to conduct subgroup analysis to assess the impact of compliance on costs, but this was not possible as implementation costs were provided in aggregate form only and could not, therefore, be linked to class-level compliance data.

Patient and public involvement

Patient and public involvement in the study was led by Common Room, a consultancy organisation whose aim is to provide a voice for children, young people and their families so that they can influence health policy, service provision and research. The director of Common Room and a team of six young research advisors made the following contributions to the project:

- Attendance at and contribution to TSC meetings. This included, for example, a standing item in which Common Room representatives reported on the various PPI activities noted below.
- Input and feedback on the presentation of child self-report surveys. This involved, for example, advice on how to optimise the survey layout to improve its overall accessibility and make survey completion an attractive proposition for participants.
- Input and feedback on standardised survey instructions, debriefs and activities for children who completed surveys early (or had been opted out by their parents) during data collection visits to schools. This included, for example, advice on how best to explain concepts such as anonymity and confidentiality to primary-school-aged children so that they truly understood what was being asked of them.
- Focus groups in schools to explore the experiences of children who had taken part in the GBG, in which they were asked to reflect on the game, the rules, how much they remembered and could describe, and whether they would recommend it to other classes and schools.

- Input and feedback on dissemination of the EEF trial findings to schools (i.e. a poster for children and a findings document for staff). Common Room supported the design process and helped to ensure a balanced approach in which stakeholder views from the IPE, as well as the substantive trial outcomes, were included.
- Input and feedback on the *Plain English summary* of this report. This key section of the report required a balance between accessibility and transparent reporting of findings.
- Development of a short film on YouTube (YouTube, LLC, San Bruno, CA, USA) to present project findings in an accessible manner to non-academic audiences.[97]
- Input and feedback on dissemination of the NIHR trial findings to schools (in process at the time of writing).

Chapter 3 Results

Parts of this chapter are reproduced or adapted with permission from the GBG trial protocol (available from the NIHR project web page: www.journalslibrary.nihr.ac.uk/programmes/phr/145238).

This chapter presents descriptive data on the trial sample, followed by analyses pertaining to hypotheses 1 to 6.

School and child characteristics

Table 3 provides summary data pertaining to school and child characteristics, alongside national averages drawn from Department for Education statistical releases in the baseline year of the trial.[98–102]

The composition of trial schools was analogous to that of primary schools in England with respect to size and the proportion of students speaking EAL. However, trial schools had significantly larger proportions of children with SEND and eligible for FSMs, and were characterised by lower rates of school absence and attainment (children achieving ≥ level 4 in English and Maths). Eight of the 77 schools (10%) were rated as 'outstanding' by Ofsted, 54 (70%) were rated as 'good', nine (12%) were rated as 'requires improvement', and six (8%) were rated as 'inadequate' in their most recent inspection prior to the commencement of the trial. The trial arms were well balanced at the school level, with no substantive differences between the intervention and control arms for these characteristics. Similarly, the differences between trial arms in terms of child-level characteristics and outcomes (*Table 4*) were negligible.[1] Therefore, the balance on key observables was considered to be good.

TABLE 3 School and child characteristics in the GBG trial and national averages

Variable	GBG (schools, *N* = 38)		Usual practice (schools, *N* = 39)		National average (2014/15)
	n (missing)	Mean (SD)	*n* (missing)	Mean (SD)	
School level (continuous)					
Size (full-time students) (*n*)	38 (0)	298.21 (134.33)	39 (0)	315.41 (186.65)	269
School absence (half-days absent) (%)	38 (0)	4.26 (0.90)	39 (0)	4.17 (0.96)	4.6
FSMs (children) (%)	38 (0)	27.56 (13.37)	39 (0)	24.46 (13.30)	15.6
EAL (children) (%)	38 (0)	22.01 (26.05)	39 (0)	23.19 (27.91)	19.4
SEND (children) (%)	38 (0)	20.85 (9.30)	39 (0)	18.17 (5.94)	15.4
Attainment (children achieving ≥ level 4 in English and Maths) (%)	38 (0)	76.21 (12.05)	39 (0)	74.87 (10.96)	80
	GBG (children, *N* = 1560)		**Usual practice (children, *N* = 1524)**		**National average (2014/15)**
	n (missing)	Percentage	*n* (missing)	Percentage	
Child level (categorical) (%)					
Sex (male)	1560 (0)	50.4	1524 (0)	54.9	50
FSMs (children)	1544 (16)	27.4	1492 (32)	22.8	15.6
EAL (children)	1544 (16)	26.1	1492 (32)	29.5	19.4
SEND (children)	1544 (16)	23.1	1492 (32)	18	15.4
SDQ conduct problems (at risk)	1499 (61)	17.9	1470 (54)	14.3	13.2

National data taken from Department for Education statistical releases.[98–102]

TABLE 4 Descriptive statistics for the GBG trial

Variable	Baseline (T1)		Follow-up (T3)		12-month follow-up (T4)		24-month follow-up (T5)	
	GBG	Usual practice	GBG	Usual practice	GBG	Usual practice	GBG	Usual practice
Conduct problems, percentage at risk (n)	18.15 (272)	14.48 (213)	13.05 (157)	12.6 (165)	15.69 (191)	13.12 (161)	13.14 (141)	13.63 (160)
Psychological well-being (range 0–100), mean (SD)	–	–	48.617 (10.574)	49.209 (10.098)	48.712 (9.862)	49.313 (9.875)	48.504 (9.238)	49.598 (9.489)
Emotional symptoms, percentage at risk (n)	–	–	10.22 (123)	10.99 (144)	12.08 (147)	11.74 (144)	13.05 (140)	13.55 (159)
Peer and social support (range 0–100), mean (SD)	–	–	51.721 (11.833)	51.699 (11.236)	52.46 (10.882)	53.177 (10.72)	50.891 (10.415)	52.95 (9.939)
School environment (range 0–100), mean (SD)	–	–	53.208 (10.902)	53.312 (10.657)	51.835 (10.25)	51.915 (10.538)	51.341 (9.234)	52.565 (9.776)
School absence, percentage of possible sessions (SD)	4.59 (4.71)	4.38 (4.98)	4.31 (5.16)	4.27 (5.27)	4.83 (6.46)	4.25 (4.95)	4.30 (5.75)	4.17 (5.20)
Bullying (i.e. social acceptance) (range 0–100), mean (SD)	–	–	45.932 (11.755)	46.009 (12.183)	47.14 (11.342)	47.821 (11.638)	49.111 (10.446)	49.067 (11.044)
Exclusion from school, no sessions excluded (SD)	SUPP	SUPP	SUPP	SUPP	0.029 (0.329)	0.036 (0.405)	0.023 (0.242)	0.046 (0.366)
Reading attainment, mean (SD)	15.108 (3.311)	15.423 (3.386)	32.489 (10.308)	33.05 (10.414)	37.258 (9.956)	37.776 (9.457)	101.503 (9.586)	102.210 (9.285)
Concentration problems (range 1–6), mean (SD)	2.602 (1.13)	2.548 (1.146)	2.548 (1.133)	2.495 (1.129)	2.437 (1.178)	2.432 (1.135)	2.352 (1.148)	2.392 (1.174)
Disruptive behaviour (range 1–6), mean (SD)	1.709 (0.81)	1.612 (0.812)	1.74 (0.856)	1.647 (0.837)	1.747 (0.854)	1.706 (0.789)	1.732 (0.84)	1.74 (0.863)
Prosocial behaviour (range 1–6), mean (SD)	4.893 (0.875)	4.946 (0.917)	4.808 (0.93)	4.932 (0.952)	4.915 (0.96)	4.917 (0.963)	4.916 (0.953)	4.842 (0.981)

SUPP, suppressed because counts were below the reporting threshold of 10 set by the ONS.

Notes
Exclusion from school statistics for the 12-month follow-up (T4) are collapsed together with exclusion from school statistics for the immediate post-intervention follow-up (T3) to avoid low counts. Reading attainment was measured via KS1 assessments at baseline (T1), the HGRT at post-intervention follow-up and 12-month follow-up (T3 and T4), and KS2 assessments at 24-month follow-up (T5).

Missing data

Data for the primary outcome, SDQ conduct problems at T3, were missing for 571 (18.5%) of the trial sample. A logistic regression (*Table 5*) identified that children with missing data had lower KS1 scores (β –0.175; $p < 0.001$), lower prosocial behaviour scores (β –0.149; $p = 0.002$) and fewer concentration problems (β –0.19; $p < 0.001$) than those without missing data. In addition, there were significant differences according to the proportion of pupils eligible for FSMs (β –0.246; $p = 0.011$) and school size (β –0.387; $p < 0.001$). We found no evidence of differences based on sex, FSM eligibility, conduct problems or disruptive behaviour. More importantly, there appeared to be no strong evidence for the trial group itself to be a significant missingness mechanism. Accordingly, as there were missing data for > 5% of the sample and it was unlikely to be missing completely at random, FIML was utilised for subsequent analyses (where possible).

Objective 1: to determine the impact of the Good Behaviour Game on health- and education-related outcomes for children

Hypothesis 1

Children in primary schools implementing the GBG over a 2-year period will demonstrate significantly better outcomes in mental health; conduct problems (hypothesis 1a), psychological well-being (hypothesis 1b) and emotional symptoms (hypothesis 1c); sources of resilience; peer and social support (hypothesis 1d) and school environment (hypothesis 1e); school absence (hypothesis 1f), and significantly lower rates of bullying (i.e. social acceptance; hypothesis 1g) and exclusion from school (hypothesis 1h) than children attending control schools.

Tables 6–8 present the findings of our ITT analyses. In relation to the primary trial outcome, conduct problems, there is no evidence of the impact of the GBG (standardised coefficient –0.039, SE 0.323; $p = 0.903$). There were also no intervention effects identified in relation to psychological well-being (standardised coefficient –0.251, SE = 0.354; $p = 0.477$) or emotional symptoms (standardised coefficient –0.265, SE = 0.318; $p = 0.405$); peer and social support (standardised coefficient –0.016, SE = 0.356; $p = 0.965$) or school environment (standardised coefficient 0.019, SE = 0.319; $p = 0.952$); school absence

TABLE 5 Multilevel binary logistic regression for missingness on SDQ conduct problems at follow-up

Conduct problems (*n* = 3084)	Standardised co-efficient β (SE)	*p*-value
Threshold (SE)	–3.720 (0.831)	
School level		
Size	–0.387 (0.090)	< 0.001
FSMs (%)	–0.246 (0.097)	0.011
Trial group (if GBG)	–0.070 (0.211)	0.739
Child level		
Sex (if male)	–0.021 (0.056)	0.706
FSMs (if eligible)	0.027 (0.028)	0.340
KS1 reading attainment	–0.175 (0.033)	< 0.001
Conduct problems T1	0.047 (0.074)	0.522
Concentration problems T1	–0.190 (0.048)	< 0.001
Disruptive behaviour T1	0.026 (0.084)	0.762
Prosocial behaviour T1	–0.149 (0.048)	0.002

TABLE 6 Multilevel binary logistic regression models of the impact of the GBG on conduct problems and emotional symptoms

Variable	Conduct problems (hypothesis 1a) (n = 3084)			Emotional symptoms (hypothesis 1c) (n = 3084)		
Threshold (SE)	3.391 (1.149)			2.248 (0.437)		
	Standardised coefficient (SE)	*p*-value	OR (95% CI)	Standardised coefficient (SE)	*p*-value	OR (95% CI)
School level						
Size	−0.137 (0.247)	0.579		−0.108 (0.155)	0.487	
FSMs (%)	−0.026 (0.240)	0.913		0.091 (0.171)	0.592	
Trial group (if GBG)	−0.039 (0.323)	0.903	0.962 (0.511 to 1.811)	−0.265 (0.318)	0.405	0.767 (0.411 to 1.431)
Variance	1.124 (0.310)			0.531 (0.165)		
Child level						
FSMs (if eligible)	0.092 (0.034)	0.006		0.084 (0.037)	0.023	
Sex (if male)	0.171 (0.036)	< 0.001		−0.100 (0.038)	0.008	
Baseline conduct problems	0.536 (0.026)	< 0.001		0.171 (0.033)	<0.001	

OR, odds ratio.

Note
Shading indicates the main study finding (impact of the GBG). M*plus* 8.4 was used for these models.

(coefficient −0.065, SE = 0.046; *p* = 0.154), bullying (i.e. social acceptance; standardised coefficient 0.191, SE = 0.393; *p* = 0.627) or exclusion from school (coefficient −0.991, SE = 0.586; *p* = 0.091). In sum, our ITT findings at immediate post-intervention follow-up (T3) were null for all outcomes.

Objective 2: to determine the impact of the Good Behaviour Game on a variety of outcomes for boys at risk of developing conduct problems

Hypothesis 2

Boys at risk of developing conduct disorders (defined as scoring in the borderline or abnormal ranges of the conduct problems subscale of the teacher-rated SDQ at baseline) in primary schools implementing the GBG over a 2-year period will demonstrate significantly better outcomes in mental health; conduct problems (hypothesis 2a), psychological well-being (hypothesis 2b) and emotional symptoms (hypothesis 2c); sources of resilience; peer and social support (hypothesis 2d) and school environment (hypothesis 2e); school absence (hypothesis 2f), and significantly lower rates of bullying (i.e. social acceptance; hypothesis 2g) and exclusion from school (hypothesis 2h) than those at-risk boys attending control schools.

Tables 9–11 present the findings of our subgroup analyses. In relation to conduct problems, there is no evidence of the impact of the GBG for at-risk boys (standardised coefficient 0.012, SE = 0.081; *p* = 0.879). There were also no intervention effects identified in relation to psychological well-being (standardised coefficient −0.039, SE = 0.507; *p* = 0.507) or emotional symptoms (standardised coefficient −0.093, SE = 0.084; *p* = 0.264); peer and social support (standardised coefficient 0.019, SE = 0.060; *p* = 0.758) or school environment (standardised coefficient −0.012, SE = 0.060; *p* = 0.849); school absence (coefficient −0.275, SE = 0.187; *p* = 0.141) or exclusion from school (coefficient 1.259, SE = 1.874; *p* = 0.502) for this subgroup. The analysis did reveal an intervention effect on bullying (i.e. social acceptance) for at-risk boys (standardised coefficient −0.125, SE = 0.051; *p* = 0.014). However, the nature of the Kidscreen-27 scoring is such that this represents a significant increase in bullying for this subgroup in GBG schools.

TABLE 7 Multilevel linear regression models of the impact of the GBG on psychological well-being, peer and social support, school environment and bullying (i.e. social acceptance)

	Psychological well-being (hypothesis 1b) ($n = 3084$)			Peer and social support (hypothesis 1d) ($n = 3084$)			School environment (hypothesis 1e) ($n = 3084$)			Bullying (i.e. social acceptance; hypothesis 1g) ($n = 3084$)		
Intercept (SE)	50.812 (0.859)			52.005 (0.951)			55.868 (0.944)			47.911 (0.885)		
	Standardised coefficient (SE)	p-value	ES (95% CI)	Standardised coefficient (SE)	p-value	ES (95% CI)	Standardised coefficient (SE)	p-value	ES (95% CI)	Standardised coefficient (SE)	p-value	ES (95% CI)
School level												
Size	–0.223 (0.166)	0.179		–0.110 (0.164)	0.503		–0.167 (0.149)	0.264		–0.397 (0.185)	0.031	
FSMs (%)	0.011 (0.185)	0.951		0.252 (0.186)	0.177		0.150 (0.165)	0.363		–0.333 (0.198)	0.092	
Trial group (if GBG)	–0.251 (0.354)	0.477	–0.038 (–0.109 to 0.032)	–0.016 (0.356)	0.965	–0.002 (–0.073 to 0.068)	0.019 (0.319)	0.952	0.004 (–0.067 to 0.074)	0.191 (0.393)	0.627	0.023 (–0.048 to 0.094)
Variance	2.287 (0.998)			2.716 (1.153)			3.491 (1.180)			1.437 (0.945)		
Child level												
FSMs (if eligible)	–0.035 (0.022)	0.003		0.006 (0.022)	0.773		0.024 (0.021)	0.260		–0.059 (0.021)	0.006	
Sex (if male)	–0.006 (0.021)	0.772		–0.035 (0.021)	0.093		–0.127 (0.020)	< 0.001		0.088 (0.020)	< 0.001	
Baseline conduct problems	–0.134 (0.022)	< 0.001		–0.058 (0.022)	0.010		–0.224 (0.021)	< 0.001		–0.099 (0.022)	< 0.001	
Variance	102.554 (2.993)			129.444 (3.738)			104.644 (3.035)			138.468 (4.004)		

Shading indicates the main study finding (impact of the GBG). M*plus* 8.4 was used for these models.

TABLE 8 Multilevel negative binomial regression models of the impact of the GBG on school absence and exclusion from school

Variable	School absence (hypothesis 1f) (n = 3007)			Exclusion from school (hypothesis 1h) (n = 3035)		
Intercept (SE)	−3.844 (0.066)			−8.476 (1.190)		
	Coefficient (SE)	p-value	IRR (95% CI)	Coefficient (SE)	p-value	IRR (95% CI)
School level						
Size	0.00006 (0.00009)	0.551		−0.002 (0.002)	0.395	
FSMs (%)	0.003 (0.002)	0.111		0.070 (0.021)	0.001	
Trial group (if GBG)	−0.065 (0.046)	0.154	0.937 (0.856 to 1.025)	−0.991 (0.586)	0.091	0.371 (0.118 to 0.855)
Variance	0.014 (0.006)			1.784 (0.917)		
Child level						
Sex (if male)	0.068 (0.030)	0.026		2.879 (0.664)	< 0.001	
FSMs (if eligible)	0.099 (0.040)	0.013		0.726 (0.429)	0.091	
Baseline (T1)	0.028 (0.002)	< 0.001		2.504 (1.479)	0.090	
Overdispersion	−0.280 (0.051)	< 0.001		2.856 (0.382)	< 0.001	

Shading indicates the main study finding (impact of the GBG). Stata, version 16.1, was used for these models.

TABLE 9 Multilevel binary logistic regression models of the impact of the GBG on conduct problems and emotional symptoms among boys at risk of developing conduct problems

Variable	Conduct problems (hypothesis 2a) (n = 3084)			Emotional symptoms (hypothesis 2c) (n = 3084)		
Threshold (SE)	3.053 (0.375)			2.266 (0.255)		
	Standardised coefficient (SE)	p-value	OR (95% CI)	Standardised coefficient (SE)	p-value	OR (95% CI)
School level						
Size	−0.182 (0.134)	0.175		−0.112 (0.116)	0.331	
FSMs (%)	0.031 (0.141)	0.826		0.118 (0.138)	0.392	
Trial group (if GBG)	−0.340 (0.354)	0.338		−0.212 (0.370)	0.566	
Variance	0.994 (0.272)			0.535 (0.152)		
Child level						
FSMs (if eligible)	0.107 (0.033)	0.001		0.085 (0.035)	0.016	
Sex (if male)	0.144 (0.058)	0.014		−0.083 (0.056)	0.136	
Baseline conduct problems (if at risk)	0.475 (0.059)	< 0.001		0.231 (0.070)	0.001	
Cross-level interactions						
Trial group*risk	−0.052 (0.084)	0.539		−0.118 (0.081)	0.145	
Trial group*sex	0.136 (0.084)	0.103		0.003 (0.073)	0.966	
Risk*sex	−0.017 (0.075)	0.817		−0.076 (0.078)	0.329	
Trial group*sex*risk	0.012 (0.081)	0.879	1.012 (0.864 to 1.186)	0.093 (0.084)	0.264	1.097 (0.931 to 1.294)

OR, odds ratio.

Note
Shading indicates the main study finding (impact of the GBG). M*plus* 8.4 was used for these models.

TABLE 10 Multilevel linear regression models of the impact of the GBG on psychological well-being, peer and social support, school environment and bullying (i.e. social acceptance) among boys at risk of developing conduct problems

Variable	Psychological well-being (hypothesis 2b) (n = 3084)			Peer and social support (hypothesis 2d) (n = 3084)			School environment (hypothesis 2e) (n = 3084)			Bullying (i.e. social acceptance; hypothesis 2g) (n = 3084)		
Intercept (SE)	50.513 (0.912)			51.373 (0.994)			55.740 (0.982)			47.773 (0.950)		
	Standardised coefficient (SE)	*p*-value	ES (95% CI)	Standardised coefficient (SE)	*p*-value	ES (95% CI)	Standardised coefficient (SE)	*p*-value	ES (95% CI)	Standardised coefficient (SE)	*p*-value	ES (95% CI)
School level												
Size	−0.186 (0.164)	0.257		−0.001 (0.002)	0.594		−0.137 (0.150)	0.364		−0.352 (0.180)	0.050	
FSMs (%)	0.007 (0.183)	0.970		0.034 (0.024)	0.154		0.139 (0.167)	0.406		−0.317 (0.195)	0.104	
Trial group (if GBG)	−0.258 (0.452)	0.567		0.520 (0.427)	0.223		−0.233 (0.407)	0.566		0.199 (0.523)	0.703	
Variance	2.448 (1.032)			2.681 (1.150)			3.419 (1.179)			1.663 (0.995)		
Child level												
Sex (if male)	−0.008 (0.030)	0.800		−0.011 (0.030)	0.710		−0.160 (0.028)	< 0.001		0.079 (0.029)	0.006	
FSMs (if eligible)	−0.040 (0.022)	0.066		0.003 (0.022)	0.896		0.017 (0.021)	0.406		−0.059 (0.020)	0.003	
Baseline conduct problems (if at risk)	−0.151 (0.059)	0.011		−0.057 (0.062)	0.354		−0.217 (0.059)	< 0.001		−0.256 (0.048)	< 0.001	
Variance	102.995 (3.007)			129.391 (3.737)			106.021 (3.075)			137.593 (3.992)		
Cross-level interactions												
Trial group*risk	0.066 (0.059)	0.270		−0.027 (0.061)	0.657		0.027 (0.061)	0.653		0.161 (0.050)	0.001	
Trial group*sex	−0.013 (0.038)	0.728		−0.066 (0.038)	0.082		0.029 (0.037)	0.427		−0.023 (0.035)	0.515	
Risk*sex	0.033 (0.061)	0.587		0.037 (0.062)	0.552		0.031 (0.062)	0.613		0.167 (0.051)	0.001	
Trial group*sex*risk	−0.039 (0.059)	0.507	−0.165 (−0.318 to −0.012)	0.019 (0.060)	0.758	0.077 (−0.075 to 0.230)	−0.012 (0.060)	0.849	−0.049 (−0.202 to 0.103)	−0.125 (0.051)	0.014	−0.563 (−0.716 to −0.409)

Shading indicates the main study finding (impact of the GBG). M*plus* 8.4 was used for these models.

TABLE 11 Multilevel negative binomial regression models of the impact of the GBG on school absence and exclusion from school among boys at risk of developing conduct problems

Variable	School absence (hypothesis 2f) (n = 2914)			Exclusion from school (hypothesis 2h) (n = 2940)		
Intercept (SE)	−3.857 (0.067)			−8.432 (1.419)		
	Coefficient (SE)	p-value	IRR (95% CI)	Coefficient (SE)	p-value	IRR (95% CI)
School level						
School size	0.00003 (0.0001)	0.676		−0.001 (0.001)	0.333	
Percentage eligible for FSMs	0.002 (0.002)	0.209		0.049 (0.020)	0.015	
Trial group (if GBG)	−0.012 (0.051)	0.811		0.383 (1.309)	0.770	
Variance	0.014 (0.006)			2.126 (0.851)		
Child level						
Sex (if male)	0.123 (0.047)	0.009		2.307 (1.157)	0.046	
FSMs (if eligible)	0.103 (0.041)	0.012		0.740 (0.478)	0.121	
Baseline	0.028 (0.002)	< 0.001		0.871 (1.041)	0.403	
At risk (conduct)	−0.262 (0.132)	0.047		3.962 (1.476)	0.007	
Overdispersion	−0.296 (0.052)	< 0.001		2.221 (0.532)	< 0.001	
Cross-level interactions						
Trial group*risk	0.317 (0.147)	0.031		−1.888 (1.846)	0.306	
Trial group*sex	−0.165 (0.063)	0.009		−1.012 (1.502)	0.500	
Risk group*sex	0.322 (0.157)	0.040		−1.601 (1.317)	0.224	
Trial group*sex*risk	−0.275 (0.187)	0.141	0.759 (0.527 to 1.095)	1.259 (1.874)	0.502	3.524 (0.089 to 138.75)

Shading indicates the main study finding (impact of the GBG). Stata, version 16.1, was used for these models. Exclusion from school at follow-up (2016/17) cannot be reported because of counts being below the minimum threshold set by the ONS of 10 pupils. Results reported here refer to exclusion from school at the 12-month follow-up (T4).

This is because the Kidscreen-27 technically labels this domain 'social acceptance', with higher scores indicative of increased social acceptance, and lower scores indicative of increased bullying.[70] In sum, our subgroup analysis findings at immediate post-intervention follow-up (T3) were null for all outcomes, with the exception of an unexpected negative effect on bullying.

Objective 3: to determine the extent to which the effects of the Good Behaviour Game vary by intervention compliance (dosage)

Hypothesis 3

The magnitude of intervention effects noted in hypotheses 1a–h above will vary as a function of intervention compliance. Specifically, we predict larger ESs in schools defined as compliers than non-compliers in terms of dosage (hypotheses 3a–h).

Classroom-level dosage across the span of the intervention period (T1–T3) ranged from 0 to 3535 minutes (mean 1066, SD 719.50). In terms of frequency and duration, teachers played the game approximately twice per week between T1 and T2, and between once and twice per week between T2 and T3; the average game session length in both years was approximately 15 minutes. Nine schools formally ceased implementation prior to T3, although their dosage data are included in

the above estimates. For the main analysis, moderate compliers (50th percentile) were identified as the classrooms in which the game was played for > 1030 minutes ($n_{student}$ = 672; 43.1%). In a sensitivity analysis, classrooms that fell above the 75th percentile (> 1348 minutes) were deemed to be high compliers ($n_{student}$ = 333; 21.3%). All CACE models were shown to have appropriate and easily distinguished classes[103,104] with no less than 1% total count, high posterior probabilities (> 90%) and acceptable entropy values (0.68–0.86).

Compliance predictors

The compliance predictors for moderate- and high-compliance models can be found in *Tables 12* and *13*. Sex was shown to be a statistically significant predictor in all moderate-compliance models, except for bullying (i.e. social acceptance) and school absence. Similarly, both concentration problems and prosocial behaviour were shown to predict compliance in all moderate models, except for school absence. Thus, classrooms with more boys and those that had lower levels of concentration problems and prosocial behaviour were more likely to comply with the intervention. All school-level characteristics were shown to predict compliance to some extent, although this varied by model (see *Table 12*).

With respect to high-compliance models, concentration problems and prosocial behaviour were shown to predict compliance, except for school absence, for which concentration problems and baseline school absence were the only significant predictors. As with the moderate-compliance models, classrooms with students that had lower levels of concentration problems and prosocial behaviour were more likely to comply with the intervention; this was true for all variables except for school absence. Only school-level EAL was found to be a statistically significant school-level predictor, and only for the conduct problems model; thus, classrooms in schools with higher percentages of students with EAL were more likely to be high compliers (see *Table 13*).

Complier-average causal effect models

The moderate- and high-compliance intervention effects for trial outcomes are summarised in *Table 14*. For full details of the models, see *Appendix 1, Tables 28–34*.

In relation to the primary trial outcome, conduct problems (hypothesis 3a), there is no evidence of intervention effects in the context of either moderate (standardised coefficient 0.006, SE 0.248; $p = 0.982$) or high compliance (standardised coefficient 0.258, SE 0.539; $p = 0.632$). Similarly, there were no intervention effects in either compliance context for emotional symptoms (hypothesis 3c: moderate compliance, standardised coefficient –0.247, SE = 0.38; $p = 0.515$; high compliance, standardised coefficient –0.341, SE 0.428; $p = 0.426$); peer and social support (hypothesis 3d: moderate compliance: standardised coefficient 0.491, SE 0.341; $p = 0.150$; high compliance, standardised coefficient –0.246, SE 2.341; $p = 0.916$) or school environment (hypothesis 3e: moderate compliance, standardised coefficient –0.121, SE 0.671; $p = 0.857$; high compliance, standardised coefficient 0.044, SE 1.224; $p = 0.972$); or bullying (i.e. social acceptance; hypothesis 3g: moderate compliance, standardised coefficient 0.37, SE 2.989; $p = 0.901$; high compliance, standardised coefficient 0.235, SE 0.592; $p = 0.692$).

In the case of psychological well-being (hypothesis 3b), intervention effects were found for both moderate (standardised coefficient –1.239, SE 0.377; $p = 0.001$) and high (standardised coefficient –0.959, SE 0.380; $p = 0.013$) compliance. However, contrary to predictions, this indicates that increased intervention compliance led to reduced psychological well-being. In terms of ES, moderate compliance led to a reduction of approximately one-quarter of a SD at T3 (ES –0.241, 95% CI –0.303 to –0.179); the high compliance effect was smaller, being equivalent to a reduction of approximately one-twelfth of a SD at T3 (ES –0.084, 95% CI –0.155 to –0.013).

In the case of school absence (hypothesis 3f), significant intervention effects were found in the context of both moderate (coefficient –0.656, SE 0.072; $p < 0.001$) and high compliance (coefficient –0.674, SE 0.162, $p < 0.001$). In line with predictions, this indicates that increased intervention compliance led to reduced absence from school. In terms of ES, children in moderate-compliance GBG classrooms

TABLE 12 Predictors of moderate compliance in the GBG trial

Compliance predictors	Outcome variables, standardised coefficient (SE), with OR (95% CI) for significant child-level predictors						
	Conduct problems (hypothesis 3a) (N = 2655)	Psychological well-being (hypothesis 3b) (N = 2623)	Emotional symptoms (hypothesis 3c) (N = 2655)	Peer and social support (hypothesis 3d) (N = 2642)	School environment (hypothesis 3e) (N = 2636)	Bullying (i.e. social acceptance; hypothesis 3g) (N = 2639)	School absence (hypothesis 3f)[a] (N = 2888)
Child level							
FSMs (if eligible)	0.243 (0.351)	0.210 (0.342)	0.233 (0.349)	0.225 (0.338)	0.175 (0.327)	0.253 (0.460)	−0.172 (0.105)
SEND (if SEND)	−0.330 (0.399)	−0.371 (0.399)	−0.311 (0.405)	−0.313 (0.382)	−0.365 (0.373)	−0.382 (0.496)	−0.062 (0.178)
KS1 attainment	−0.121 (0.586)	−0.183 (0.592)	−0.127 (0.584)	−0.089 (0.576)	−0.188 (0.587)	−0.136 (0.852)	0.011 (0.027)
Sex (if male)	0.302 (0.124)*	0.267 (0.125)*	0.296 (0.123)*	0.268 (0.115)*	0.308 (0.128)*	0.283 (0.193)	0.151 (0.096)
	OR 1.352 (1.060 to 1.725)	OR 1.307 (1.022 to 1.670)	OR 1.345 (1.057 to 1.711)	OR 1.307 (1.043 to 1.639)	OR 1.361 (1.060 to 1.748)		
Concentration problems T1	−0.258 (0.070)***	−0.246 (0.062)***	−0.265 (0.071)***	−0.242 (0.055)***	−0.275 (0.083)**	−0.248 (0.112)*	−0.055 (0.083)
	OR 0.772 (0.674 to 0.886)	OR 0.987 (0.692 to 0.884)	OR 0.767 (0.668 to 0.881)	OR 0.785 (0.779 to 1.232)	OR 0.760 (0.646 to 0.894)	OR .780 (0.626 to 0.972)	
Conduct problems T1	−0.009 (0.119)	−0.013 (0.126)	−0.011 (0.119)	−0.021 (0.117)	−0.022 (0.112)	−0.009 (0.137)	0.022 (0.042)
Prosocial behaviour T1	−0.761 (0.368)*	−0.750 (0.352)*	−0.779 (0.366)*	−0.752 (0.330)*	−0.834 (0.354)*	−0.803 (0.364)*	−0.073 (0.117)
	OR 0.467 (0.227 to 0.960)	OR 0.472 (0.237 to 0.941)	OR 0.459 (0.224 to 0.941)	OR 0.471 (0.247 to 0.900)	OR 0.434 (0.218 to 0.869)	OR 0.448 (0.220 to 0.914)	
School absence T1	N/A	N/A	N/A	N/A	N/A	N/A	4.854 (1.541)***
							OR 128.19 (6.25 to 2628.47)

	Outcome variables, standardised coefficient (SE), with OR (95% CI) for significant child-level predictors						
Compliance predictors	Conduct problems (hypothesis 3a) (N = 2655)	Psychological well-being (hypothesis 3b) (N = 2623)	Emotional symptoms (hypothesis 3c) (N = 2655)	Peer and social support (hypothesis 3d) (N = 2642)	School environment (hypothesis 3e) (N = 2636)	Bullying (i.e. social acceptance; hypothesis 3g) (N = 2639)	School absence (hypothesis 3f)[a] (N = 2888)
School level							
School size	−1.589 (0.459)**	3.177 (2.173)	3.809 (2.082)	2.899 (4.308)	3.867 (3.271)	−1.131 (2.424)	−0.001 (0.001)
FSMs (%)	−0.330 (1.370)	−2.483 (2.281)	−4.385 (1.944)*	−1.811 (2.885)	−1.214 (2.514)	−2.509 (13.118)	−0.002 (0.012)
Conduct problems T1	4.028 (1.862)*	0.199 (5.521)	7.133 (4.368)	3.455 (8.644)	2.456 (5.206)	2.089 (17.141)	−0.214 (0.333)
EAL (%)	4.514 (1.455)**	−1.174 (1.178)	2.006 (0.988)*	0.087 (1.354)	0.011 (1.436)	0.064 (18.609)	−0.0002 (0.006)
KS1 attainment	−1.263 (0.873)	−2.662 (1.244)*	−2.589 (1.271)*	−2.083 (2.270)	−2.998 (1.666)	−1.606 (2.176)	0.003 (0.016)

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.
N/A, not applicable; OR, odds ratio.
a Robust SEs, clustered by school.

Note
M*plus* 8.4 was used for these models, except for school absence (hypothesis 3f), for which Stata, version 16.1, was used.

TABLE 13 Predictors of high compliance in the GBG trial

Compliance predictors	Outcome variables, standardised coefficient (SE), with OR (95% CI) for significant child-level predictors						
	Conduct problems (hypothesis 3a) (N = 2655)	Psychological well-being (hypothesis 3b) (N = 2623)	Emotional symptoms (hypothesis 3c) (N = 2655)	Peer and social support (hypothesis 3d) (N = 2642)	School environment (hypothesis 3e) (N = 2636)	Bullying (i.e. social acceptance; hypothesis 3g) (N = 2639)	School absence (hypothesis 3f)[a] (N = 2888)
Child level							
FSMs (if eligible)	0.378 (0.208)	0.287 (0.189)	0.378 (0.210)	0.350 (0.552)	0.360 (0.199)	0.324 (0.205)	–0.259 (0.140)
SEND (if SEND)	0.908 (0.712)	0.797 (0.630)	0.922 (0.725)	0.922 (1.658)	0.878 (0.673)	0.876 (0.678)	–0.291 (0.200)
KS1 attainment	–0.774 (0.763)	–0.844 (0.804)	–0.760 (0.755)	–0.765 (0.817)	–0.789 (0.763)	–0.771 (0.742)	0.061 (0.032)
Sex (if male)	0.139 (0.139)	0.131 (0.146)	0.134 (0.139)	0.130 (0.299)	0.128 (0.142)	0.140 (0.137)	0.019 (0.105)
Concentration problems T1	–0.656 (0.205)**	–0.594 (0.194)**	–0.657 (0.207)**	–0.678 (0.301)*	–0.646 (0.193)**	–0.655 (0.187)***	0.170 (0.080)*
	OR 0.519 (0.348 to 0.775)	OR 0.552 (0.377 to 0.808)	OR 0.518 (0.345 to 0.777)	OR 0.508 (0.281 to 0.917)	OR 0.524 (0.359 to 0.766)	OR 0.519 (0.360 to 0.749)	OR 1.185 (1.013 to 1.386)
Conduct problems T1	–0.010 (0.282)	–0.054 (0.264)	–0.011 (0.286)	–0.016 (0.313)	–0.014 (0.275)	–0.028 (0.279)	–0.028 (0.059)
Prosocial behaviour T1	–1.130 (0.166)***	–1.050 (0.162)***	–1.138 (0.162)***	–1.168 (0.000)[b]	–1.110 (0.173)***	–1.152 (0.180)***	–0.132 (0.131)
	OR 0.323 (0.233 to 0.447)	OR 0.350 (0.255 to 0.480)	OR 0.320 (0.233 to 0.440)		OR 0.330 (0.235 to 0.463)	OR 0.316 (0.222 to 0.450)	
School absence T1	N/A	N/A	N/A	N/A	N/A	N/A	6.118 (1.959)**
							OR 454.1 (9.8 to 21134.1)

Compliance predictors	Outcome variables, standardised coefficient (SE), with OR (95% CI) for significant child-level predictors						
	Conduct problems (hypothesis 3a) (N = 2655)	Psychological well-being (hypothesis 3b) (N = 2623)	Emotional symptoms (hypothesis 3c) (N = 2655)	Peer and social support (hypothesis 3d) (N = 2642)	School environment (hypothesis 3e) (N = 2636)	Bullying (i.e. social acceptance; hypothesis 3g) (N = 2639)	School absence (hypothesis 3f)[a] (N = 2888)
School level							
Size	1.460 (0.787)	0.640 (1.925)	8.220 (7.146)	0.538 (9.372)	1.079 (2.901)	1.617 (2.976)	0.001 (0.001)
FSMs (%)	−1.262 (1.394)	0.356 (2.296)	−2.047 (1.183)	1.245 (5.943)	−4.694 (4.004)	−0.544 (3.671)	−0.013 (0.014)
Conduct problems T1	−3.379 (1.845)	2.991 (6.633)	−22.694 (14.054)	8.291 (67.680)	5.530 (18.106)	−3.838 (8.802)	−0.452 (0.391)
EAL (%)	11.652 (5.051)*	−0.878 (1.158)	26.611 (14.168)	−0.657 (24.831)	−0.374 (2.138)	−1.969 (1.638)	−0.003 (0.007)
KS1 attainment	−1.115 (0.910)	0.410 (1.111)	−0.763 (1.266)	−1.713 (7.151)	1.573 (2.082)	0.514 (1.443)	−0.007 (0.018)

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.
OR, odds ratio.
a Robust SEs, clustered by school.
b Fixed value.

Note
M*plus* 8.4 was used for these models, except for school absence (hypothesis 3f), for which Stata, version 16.1, was used.

TABLE 14 Moderate- and high-compliance intervention effects in the GBG trial

| Outcomes | CACE effects | | | | | |
| | Moderate compliance | | | High compliance | | |
	Standardised coefficient (SE)	p-value	OR (95% CI)	Standardised coefficient (SE)	p-value	OR (95% CI)
Logistic regression						
Conduct problems (hypothesis 3a) (n = 2655)	0.006 (0.248)	0.982	1.006 (0.618 to 1.636)	0.258 (0.539)	0.632	1.294 (0.45 to 3.728)
Emotional symptoms (hypothesis 3c) (n = 2655)	−0.247 (0.38)	0.515	0.781 (0.371 to 1.644)	−0.341 (0.428)	0.426	0.711 (0.308 to 1.644)
	Standardised coefficient (SE)	p-value	Cohen's d (95% CI)	Standardised coefficient (SE)	p-value	Cohen's d (95% CI)
Linear regression						
Psychological well-being (hypothesis 3b) (n = 2623)	−1.239 (0.377)	0.001	−0.241 (−0.312 to −0.17)	−0.959 (0.38)	0.013	−0.294 (−0.365 to −0.223)
Peer and social support (hypothesis 3d) (n = 2642)	0.491 (0.341)	0.150	0.103 (0.032 to 0.174)	−0.246 (2.341)	0.916	−0.084 (−0.155 to −0.013)
School environment (hypothesis 3e) (n = 2636)	−0.121 (0.671)	0.857	−0.026 (−0.097 to 0.045)	0.044 (1.224)	0.972	0.011 (−0.06 to 0.082)
Bullying (i.e. social acceptance; hypothesis 3g) (n = 2639)	0.37 (2.989)	0.901	0.045 (−0.026 to 0.116)	0.235 (0.592)	0.692	0.03 (−0.041 to 0.101)
	Coefficient (SE)	p-value	IRR (95% CI)	Coefficient (SE)	p-value	IRR (95% CI)
Poisson regression						
School absence (hypothesis 3f) (n = 2888)	−0.656 (0.072)	< 0.001	0.519 (0.450 to 0.598)	−0.674 (0.162)	< 0.001	0.510 (0.371 to 0.701)

OR, odds ratio.

Note
M*plus* 8.4 was used for these models, except for the Poisson regression models, for which Stata, version 16.1, was used.

have an incidence rate of 51.9% of that of the would-be compliers in the control group (IRR 0.519, 95% CI 0.450 to 0.598); in the high-compliance classrooms, the ES was slightly larger, with an incidence rate of 51% (IRR 0.510, 95% CI 0.371 to 0.701).

In sum, our CACE analysis findings at the immediate post-intervention follow-up (T3) were null for all outcomes, with the exception of an unexpected intervention effect on psychological well-being and an expected intervention effect on school absence.

Objective 4: to determine whether the effects of the Good Behaviour Game are sustained (or emerge) over time

Hypothesis 4
The effects of the GBG on mental health; conduct problems (hypothesis 4a), psychological well-being (hypothesis 4b) and emotional symptoms (hypothesis 4c); sources of resilience; peer and social support (hypothesis 4d) and school environment (hypothesis 4e); school absence (hypothesis 4f), bullying (i.e. social acceptance; hypothesis 4g) and exclusion from school (hypothesis 4h), reading attainment (hypothesis 4i), prosocial behaviour (hypothesis 4j), concentration problems (hypothesis 4k) and disruptive behaviour (hypothesis 4l), will be maintained at 12- and 24-month post-intervention follow-ups.

Tables 15–18 present the findings of our longer-term follow-up analyses at 12 months post intervention (T4).

At the 12-month follow-up, there was no evidence of the impact of the GBG on conduct problems (hypothesis 4a: standardised coefficient −0.235, SE 0.261; $p = 0.369$). There were also no intervention effects identified in relation to psychological well-being (hypothesis 4b: standardised coefficient −0.259, SE 0.306; $p = 0.396$) or emotional symptoms (hypothesis 4c: standardised coefficient −0.228,

TABLE 15 Multilevel binary logistic regression models of the impact of the GBG on conduct problems and emotional symptoms at the 12-month follow-up

Variable	Conduct problems (hypothesis 4a) (n = 3084)			Emotional symptoms (hypothesis 4c) (n = 3084)		
Threshold (SE)	3.251 (0.000)			2.355 (0.439)		
	Standardised coefficient (SE)	p-value	OR (95% CI)	Standardised coefficient (SE)	p-value	OR (95% CI)
School level						
Size	−0.264 (0.101)	0.009		−0.007 (0.144)	0.964	
FSMs (%)	0.186 (0.114)	0.104		0.087 (0.177)	0.622	
Trial group (if GBG)	−0.235 (0.261)	0.369	0.791 (0.474 to 1.319)	−0.228 (0.292)	0.435	0.796 (0.449 to 1.411)
Variance	0.844 (0.276)			0.647 (0.211)		
Child level						
FSMs (if eligible)	0.150 (0.031)	< 0.001		0.114 (0.035)	0.001	
Sex (if male)	0.183 (0.035)	< 0.001		−0.102 (0.036)	0.005	
Baseline conduct problems	0.497 (0.027)	< 0.001		0.155 (0.034)	< 0.001	

OR, odds ratio.

Note
Shading indicates the main study finding (impact of the GBG). M*plus* 8.4 was used for these models.

TABLE 16 Multilevel linear regression models of the impact of the GBG on psychological well-being, peer and social support, school environment and bullying (i.e. social acceptance) at the 12-month follow-up

Variable	Psychological well-being (hypothesis 4b) (n = 3084)			Peer and social support (hypothesis 4d) (n = 3084)			School environment (hypothesis 4e) (n = 3084)			Bullying (i.e. social acceptance; hypothesis 4g) (n = 3084)		
Intercept (SE)	48.728 (0.878)			52.770 (0.838)			53.659 (1.012)			48.210 (1.044)		
	Standardised coefficient (SE)	p-value	ES (95% CI)	Standardised coefficient (SE)	p-value	ES (95% CI)	Standardised coefficient (SE)	p-value	ES (95% CI)	Standardised coefficient (SE)	p-value	ES (95% CI)
School level												
Size	−0.057 (0.142)	0.689		−0.028 (0.165)	0.866		−0.043 (0.137)	0.756		0.068 (0.149)	0.650	
FSMs (%)	0.414 (0.151)	0.006		0.480 (0.186)	0.010		0.270 (0.145)	0.063		−0.179 (0.166)	0.282	
Trial group (if GBG)	−0.259 (0.306)	0.396	−0.049 (−0.119 to 0.022)	−0.555 (0.363)	0.126	−0.074 (−0.145 to −0.003)	−0.014 (0.291)	0.961	−0.003 (−0.074 to 0.067)	−0.226 (0.321)	0.481	−0.041 (−0.112 to 0.029)
Variance	2.769 (0.988)			1.472 (0.831)			4.711 (1.291)			4.065 (1.412)		
Child level												
Sex (if male)	0.025 (0.021)	0.236		−0.040 (0.021)	0.057		−0.126 (0.021)	< 0.001		0.095 (0.021)	< 0.001	
FSMs (if eligible)	−0.070 (0.022)	0.001		−0.001 (0.022)	0.981		−0.069 (0.022)	0.001		−0.047 (0.022)	0.034	
Baseline conduct problems	−0.121 (0.023)	< 0.001		−0.070 (0.023)	0.002		−0.202 (0.022)	< 0.001		−0.146 (0.023)	< 0.001	
Variance	92.716 (2.743)			113.880 (3.358)			96.022 (2.839)			124.876 (3.718)		

Shading indicates the main study finding (impact of the GBG). M*plus* 8.4 was used for these models.

TABLE 17 Multilevel negative binomial regression models of the impact of the GBG on school absence and exclusion from school at the 12-month follow-up

Variable	School absence (hypothesis 4f) (n = 2980)			Exclusion from school (hypothesis 4h) (n = 3035)		
Intercept (SE)	−3.715 (0.072)			−6.754 (0.876)		
	Coefficient (SE)	p-value	IRR (95% CI)	Coefficient (SE)	p-value	IRR (95% CI)
School level						
Size	−0.0002 (0.0001)	0.284		−0.002 (0.001)	0.106	
FSMs (%)	0.002 (0.002)	0.304		0.053 (0.021)	0.009	
Trial group (if GBG)	0.061 (0.043)	0.149	1.063 (0.978 to 1.156)	−0.913 (0.575)	0.112	0.401 (0.130 to 1.238)
Variance	0.011 (0.007)			1.847 (0.976)		
Child level						
Sex (if male)	0.033 (0.041)	0.425		2.239 (0.385)	< 0.001	
FSMs (if eligible)	0.228 (0.047)	< 0.001		0.994 (0.424)	0.019	
Baseline	0.025 (0.001)	< 0.001		2.269 (0.888)	0.011	
Overdispersion	−0.165 (0.052)	0.002		3.117 (0.485)	< 0.001	

Shading indicates the main study finding (impact of the GBG). Stata, version 16.1, was used for these models.

SE 0.292; $p = 0.435$); peer and social support (hypothesis 4d: standardised coefficient −0.555, SE 0.363; $p = 0.126$) or school environment (hypothesis 4e: standardised coefficient −0.014, SE 0.291; $p = 0.961$); school absence (hypothesis 4f: coefficient 0.061, SE 0.043; $p = 0.149$), bullying (i.e. social acceptance; hypothesis 4g: standardised coefficient −0.226, SE 0.321; $p = 0.481$) or exclusion from school (hypothesis 4h: coefficient −0.913, SE 0.575; $p = 0.112$) at this time point.

Furthermore, there was no evidence of intervention effects at the 12-month follow-up in relation to reading attainment (hypothesis 4i: standardised coefficient 0.100, SE 0.229; $p = 0.661$), concentration problems (hypothesis 4k: standardised coefficient −0.221, SE 0.240; $p = 0.356$), disruptive behaviour (hypothesis 4l: standardised coefficient −0.190, SE 0.240; $p = 0.428$) or prosocial behaviour (hypothesis 4j: standardised coefficient 0.164, SE 0.242; $p = 0.499$). In sum, our ITT findings at 12-month post-intervention follow-up (T4) were null for all outcomes.

Tables 19–22 present the findings of our longer-term follow-up at 24 months post intervention (T5).

At the 24-month follow-up, there was no evidence of the impact of the GBG on conduct problems (hypothesis 4a: standardised coefficient −0.400, SE 0.273; $p = 0.143$). There were also no intervention effects identified in relation to psychological well-being (hypothesis 4b: standardised coefficient −0.389, SE 0.260; $p = 0.135$) or emotional symptoms (hypothesis 4c: standardised coefficient −0.244, SE 0.321; $p = 0.449$); school environment (hypothesis 4e: standardised coefficient −0.485, SE 0.262; $p = 0.064$); school absence (hypothesis 4f: standardised coefficient −0.014, SE 0.049; $p = 0.770$), bullying (i.e. social acceptance; hypothesis 4g: standardised coefficient 0.036, SE 0.304; $p = 0.907$) or exclusion from school (hypothesis 4h: standardised coefficient 0.703, SE 0.490; $p = 0.151$) at this time point.

Furthermore, there was no evidence of intervention effects at the 24-month follow-up in relation to reading attainment (hypothesis 4i: standardised coefficient −0.006, SE 0.059; $p = 0.919$), concentration problems (hypothesis 4k: standardised coefficient −0.171, SE 0.253; $p = 0.500$), disruptive behaviour (hypothesis 4l: standardised coefficient −0.270, SE 0.252; $p = 0.285$) or prosocial behaviour (hypothesis 4j: standardised coefficient 0.300, SE 0.245; $p = 0.220$).

TABLE 18 Multilevel linear regression models of the impact of the GBG on reading attainment, concentration problems, disruptive behaviour and prosocial behaviour at the 12-month follow-up

Variable	Reading attainment (hypothesis 4i) (n = 3084)			Concentration problems (hypothesis 4k) (n = 3084)			Disruptive behaviour (hypothesis 4l) (n = 3084)			Prosocial behaviour (hypothesis 4j) (n = 3084)		
Intercept (SE)	5.496 (1.091)			0.492 (0.149)			0.618 (0.120)			3.015 (0.183)		
	Standardised coefficient (SE)	p-value	ES (95% CI)	Standardised coefficient (SE)	p-value	ES (95% CI)	Standardised coefficient (SE)	p-value	ES (95% CI)	Standardised coefficient (SE)	p-value	ES (95% CI)
School level												
Size	−0.060 (0.110)	0.581		0.130 (0.117)	0.267		−0.037 (0.118)	0.756		−0.089 (0.119)	0.454	
FSMs (%)	−0.521 (0.100)	< 0.001		0.190 (0.119)	0.112		0.182 (0.119)	0.125		−0.110 (0.121)	0.362	
Trial group (if GBG)	0.100 (0.229)	0.661	0.036 (−0.035 to 0.107)	−0.221 (0.240)	0.356	−0.096 (−0.167 to −0.025)	−0.190 (0.240)	0.428	−0.091 (−0.162 to −0.020)	0.164 (0.242)	0.499	0.076 (0.006 to 0.147)
Variance	3.890 (0.839)			0.148 (0.029)			0.102 (0.019)			0.164 (0.031)		
Child level												
Sex (if male)	−0.025 (0.013)	0.067		0.140 (0.017)	< 0.001		0.143 (0.017)	< 0.001		−0.132 (0.018)	< 0.001	
FSMs (if eligible)	−0.017 (0.014)	0.229		0.071 (0.017)	< 0.001		0.095 (0.017)	< 0.001		−0.104 (0.019)	< 0.001	
Baseline	0.774 (0.009)	< 0.001		0.628 (0.014)	< 0.001		0.590 (0.015)	< 0.001		0.450 (0.019)	< 0.001	
Variance	37.821 (1.129)			0.692 (0.020)			0.363 (0.011)			0.604 (0.018)		

Shading indicates the main study finding (impact of the GBG). M*plus* 8.4 was used for these models.

TABLE 19 Multilevel binary logistic regression models of the impact of the GBG on conduct problems and emotional symptoms at the 24-month follow-up

Variable	Conduct problems (hypothesis 4a) (n = 3084)			Emotional symptoms (hypothesis 4c) (n = 3084)		
Threshold (SE)	4.451 (0.000)			2.065 (0.363)		
	Standardised coefficient (SE)	*p*-value	OR (95% CI)	Standardised coefficient (SE)	*p*-value	OR (95% CI)
School level						
Size	0.133 (0.116)	0.251		0.105 (0.156)	0.500	
FSMs (%)	0.340 (0.127)	0.007		–0.088 (0.198)	0.659	
Trial group (if GBG)	–0.400 (0.273)	0.143	0.670 (0.393 to 1.145)	–0.244 (0.321)	0.449	0.783 (0.418 to 1.469)
Variance	0.618 (0.217)			0.371 (0.130)		
Child level						
Sex (if male)	0.242 (0.037)	< 0.001		–0.106 (0.035)	0.003	
FSMs (if eligible)	0.136 (0.032)	< 0.001		0.108 (0.035)	0.002	
Baseline conduct problems	0.464 (0.029)	< 0.001		0.148 (0.034)	< 0.001	

OR, odds ratio.

Note
Shading indicates the main study finding (impact of the GBG). M*plus* 8.4 was used for these models.

In the case of peer and social support (hypothesis 4d), there was a significant intervention effect at the 24-month follow-up (standardised coefficient –0.743, SE 0.241; $p = 0.002$). Contrary to predictions, however, this was a negative effect; in other words, children in GBG schools reported significantly lower levels of peer and social support than their counterparts did in the control schools. The associated ES was roughly equivalent to one-fifth of a SD (ES –0.195). In sum, our ITT findings at the 24-month post-intervention follow-up (T4) were null for all outcomes, with the exception of an unexpected negative effect on peer and social support.

Objective 5: to assess the temporal association between mental health and academic attainment

Hypothesis 5
Children's educational and health-related outcomes will be related over time.

The chi-squared difference test indicated that the model in which time-invariant covariates (e.g. trial group, sex, shared risk) were freely estimated (hypothesis 1) provided a significantly better fit for our data than the model in which they were constrained to zero (hypothesis 0) (χ^2 difference 954.831, degrees of freedom 9; $p < 0.001$). Longitudinal measurement invariance tests were conducted for the measures of conduct problems and emotional symptoms over time. Change in comparative fit index (CFI) was used to determine significant changes in fit. The differences in CFI between the configural and the scalar invariance for conduct problems and emotional symptoms were < 0.01.

A preliminary model without time-invariant covariates displayed an acceptable overall fit, with a CFI of 0.95 (scaled CFI 0.90), Tucker–Lewis index (TLI) of 0.95 (scaled TLI 0.90) and root-mean-square error of approximation (RMSEA) of 0.062 (scaled RMSEA 0.055). This model is statistically equivalent to hypothesis 0. The nested modelling procedure required the specification of parameter constraints on time-invariant covariates, but this had the undesired effect of lowering the overall model fit, with a CFI of 0.77 (scaled CFI 0.59), TLI of 0.8 (scaled TLI 0.63) and RMSEA of 0.113 (scaled RMSEA 0.096).

TABLE 20 Multilevel linear regression models of the impact of the GBG on psychological well-being, peer and social support, school environment and bullying (i.e. social acceptance) at the 24-month follow-up

Variable	Psychological well-being (hypothesis 4b) (n = 3084)			Peer and social support (hypothesis 4d) (n = 3084)			School environment (hypothesis 4e) (n = 3084)			Bullying (i.e. social acceptance; hypothesis 4g) (n = 3084)		
Intercept (SE)	47.131 (1.021)			52.009 (1.061)			53.392 (1.059)			48.406 (1.068)		
	Standardised coefficient (SE)	p-value	ES (95% CI)	Standardised coefficient (SE)	p-value	ES (95% CI)	Standardised coefficient (SE)	p-value	ES (95% CI)	Standardised coefficient (SE)	p-value	ES (95% CI)
School level												
School size	0.121 (0.123)	0.326		0.298 (0.117)	0.011		0.005 (0.128)	0.968		0.028 (0.141)	0.841	
FSMs (%)	0.339 (0.131)	0.010		−0.079 (0.134)	0.556		0.198 (0.137)	0.151		0.207 (0.156)	0.185	
Trial group (if GBG)	−0.389 (0.260)	0.135	−0.106 (−0.176 to −0.035)	−0.743 (0.241)	0.002	−0.195 (−0.265 to −0.124)	−0.485 (0.262)	0.064	−0.133 (−0.203 to −0.062)	0.036 (0.304)	0.907	0.008 (−0.063 to 0.078)
Variance	5.308 (1.401)			5.312 (1.486)			5.954 (1.534)			4.986 (1.449)		
Child level												
Sex (if male)	0.085 (0.021)	< 0.001		0.004 (0.022)	0.840		−0.088 (0.021)	< 0.001		0.074 (0.021)	0.001	
FSMs (if eligible)	−0.080 (0.023)	< 0.001		−0.057 (0.023)	0.012		−0.056 (0.023)	0.013		−0.088 (0.023)	< 0.001	
Baseline conduct problems	−0.092 (0.024)	< 0.001		−0.061 (0.024)	0.013		−0.169 (0.024)	< 0.001		−0.137 (0.024)	< 0.001	
Variance	80.727 (2.453)			97.012 (2.943)			81.335 (2.472)			107.703 (3.272)		

Shading indicates the main study finding (impact of the GBG). M*plus* 8.4 was used for these models.

TABLE 21 Negative binomial regression models of the impact of the GBG on school absence and exclusion from school at 24-month follow-up

Variable	School absence (hypothesis 4f) (n = 2962)			Exclusion from school (hypothesis 4h) (n = 3037)		
Intercept (SE)	–3.908 (0.081)			–6.184 (0.770)		
	Coefficient (SE)	*p*-value	IRR (95% CI)	Coefficient (SE)	*p*-value	IRR (95% CI)
School level						
Size	–0.0001 (0.0002)	0.720		0.001 (0.001)	0.506	
FSMs (%)	0.007 (0.002)	0.000		0.020 (0.020)	0.325	
Trial group (if GBG)	–0.014 (0.049)	0.770	0.986 (0.896 to 1.085)	–0.703 (0.490)	0.151	0.495 (0.190 to 1.293)
Variance	0.011 (0.007)			1.847 (0.976)		
Child level						
Sex (if male)	0.139 (0.040)	0.000		2.099 (0.404)	0.000	
FSMs (if eligible)	0.246 (0.048)	0.000		0.889 (0.419)	0.034	
Baseline	0.023 (0.002)	0.001		0.442 (1.008)	0.661	
Overdispersion	–0.123 (0.045)	0.003		3.409 (0.263)	0.000	

Shading indicates the main study finding (impact of the GBG). Stata, version 16.1, was used for these models.

This is most likely to be due to the estimation of means and variance for the covariates, which have an effect on the estimated coefficients of both the measurement and structural parts when comparing hypothesis 0 with the preliminary model.

The overall model fit of hypothesis 1 was also poor, but, as shown by the chi-squared difference test, it was an improvement from hypothesis 0, with a CFI of 0.81 (scaled CFI 0.63), TLI of 0.83 (scaled TLI 0.66) and RMSEA of 0.087 (scaled RMSEA 0.092). Below, we report model hypothesis 1, depicted visually in *Figure 2* (between-person effects) and *Figure 3* (within-person effects). For full details of the models, see *Appendix 2, Tables 35–42*.

Inspection of *Figure 2* reveals significant between-person associations in emotional symptoms, conduct problems, and reading attainment. Thus, individuals experiencing emotional symptoms tend to also experience conduct problems. The between-person associations for both of these mental health variables and reading attainment were negative; hence, individuals experiencing emotional symptoms and/or conduct problems tend to also have lower reading scores. Furthermore, there are differential effects of our time-invariant covariates. Shared risk is significantly *positively* associated with emotional symptoms and conduct problems, and significantly *negatively* associated with reading attainment. In each case, a 1-unit increase in shared risk equates to change of approximately one-quarter of a SD in a given outcome variable. Sex is significantly *positively* associated with conduct problems and significantly *negatively* associated with emotional symptoms and reading attainment. Thus, being male is associated with an increase of just over half a SD in conduct problems, just below a fifth of a SD decrease in emotional symptoms and just over a tenth of a SD decrease in reading scores. Trial group is not significantly associated with any of these outcomes.

Figure 3 provides evidence of several notable temporal trends in the time-varying within-person part of the model. First, we observe significant increases in conduct problems over time, contrasted with significant decreases in reading scores. Second, there is evidence of significant cross-lagged effects, wherein emotional symptoms are negatively associated with later conduct problems. Thus, an increase of one SD in emotional symptoms at T3 is associated with just less than a fifth of a SD decrease in

TABLE 22 Multilevel linear regression models of the impact of the GBG on reading attainment, concentration problems, disruptive behaviour and prosocial behaviour at 24-month follow-up

Variable	Reading attainment (hypothesis 4i) (n = 2904)			Concentration problems (hypothesis 4k) (n = 3084)			Disruptive behaviour (hypothesis 4l) (n = 3084)			Prosocial behaviour (hypothesis 4j) (n = 3084)		
Intercept (SE)	0.150 (0.089)			0.670 (0.167)			0.500 (0.109)			2.803 (0.186)		
	Coefficient (SE)	p-value	ES (95% CI)	Standardised coefficient (SE)	p-value	ES (95% CI)	Standardised coefficient (SE)	p-value	ES (95% CI)	Standardised coefficient (SE)	p-value	ES (95% CI)
School level												
Size	−0.00007 (0.0002)	0.693		0.052 (0.121)	0.671		0.182 (0.119)	0.128		−0.155 (0.118)	0.187	
FSMs (%)	−0.002 (0.002)	0.382		0.030 (0.130)	0.815		0.115 (0.131)	0.378		−0.036 (0.127)	0.779	
Trial group (if GBG)	−0.006 (0.059)	0.919	−0.009 (−0.079 to 0.062)	−0.171 (0.253)	0.500	−0.076 (−0.146 to −0.005)	−0.270 (0.252)	0.285	−0.104 (−0.174 to −0.033)	0.300 (0.245)	0.220	0.136 (0.065 to 0.206)
Variance	0.054 (0.011)			0.179 (0.035)			0.070 (0.015)			0.149 (0.029)		
Child level												
Sex (if male)	−0.120 (0.024)	< 0.001		0.187 (0.018)	< 0.001		0.163 (0.018)	< 0.001		−0.108 (0.019)	< 0.001	
FSMs (if eligible)	−0.043 (0.029)	0.140		0.070 (0.018)	< 0.001		0.079 (0.019)	< 0.001		−0.100 (0.020)	< 0.001	
Baseline outcome	0.739 (0.013)	< 0.001		0.579 (0.015)	< 0.001		0.561 (0.016)	< 0.001		0.467 (0.019)	< 0.001	
Variance	0.418 (0.011)			0.753 (0.023)			0.438 (0.013)			0.621 (0.019)		

Shading indicates the main study finding (impact of the GBG). M*plus* 8.4 was used for these models, except for the reading attainment regression model, for which Stata, version 16.1, was used.
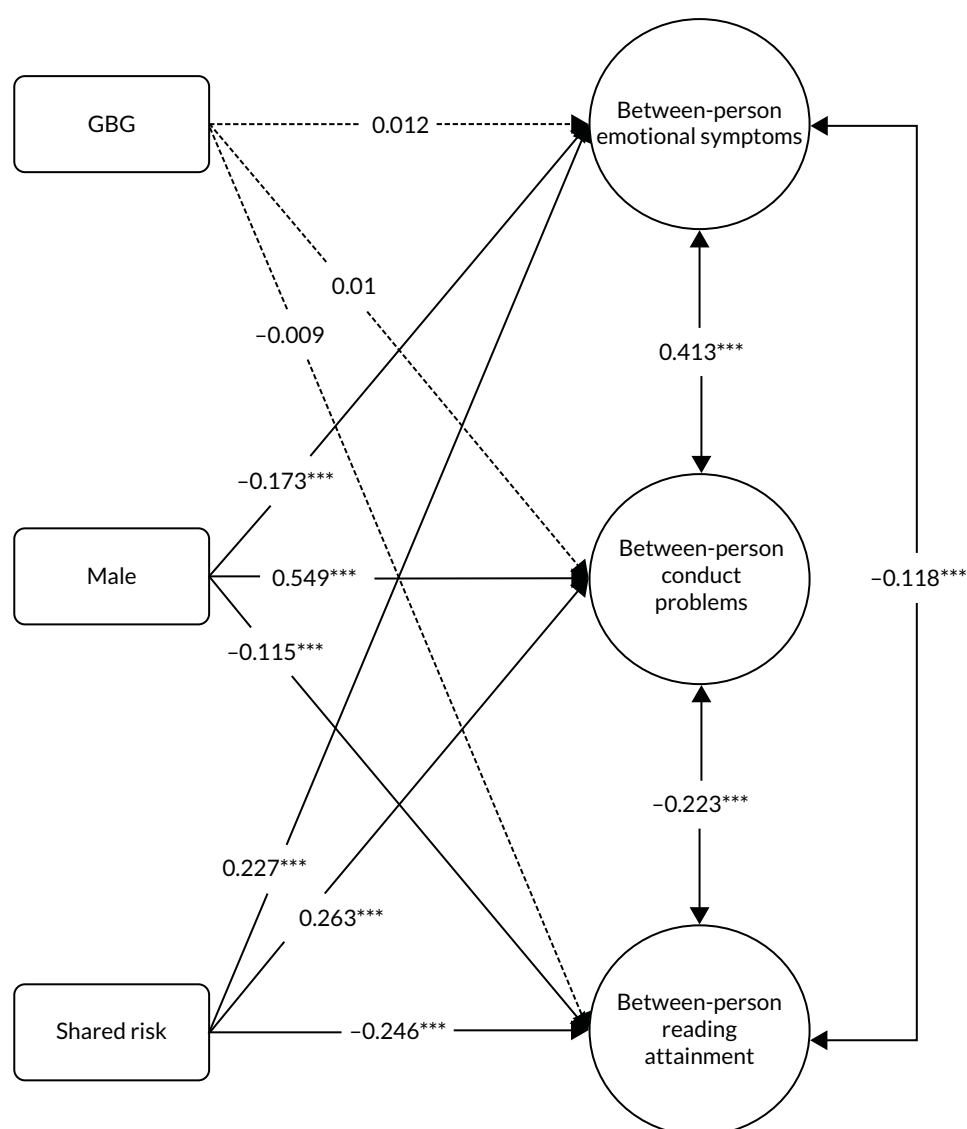
FIGURE 2 Between-person effects for emotional symptoms, conduct problems and reading attainment, and the influence of trial group, sex and shared risk (number of participants = 2987). ***$p < 0.001$; standard arrow, statistically significant pathway; dashed arrow, non-significant pathway.

conduct problems at T4. This effect seems to increase over time, with a one-SD increase in emotional symptoms at T4 being associated with approximately a quarter of a SD decrease in conduct problems at T5. However, in relation to hypothesis 5, there is no evidence of significant cross-lagged effects between either emotional symptoms or conduct problems and attainment.

Objective 6: to assess the health economic impact of the Good Behaviour Game

Hypothesis 6

The GBG will represent an efficient use of resources when considered from a public-sector perspective.

The economic analysis comprised a CCA of the available GBG data. The analysis was conducted from a public-sector perspective, with the inclusion of indirect costs to parents and caregivers. Up to 79% of the total costs of conduct disorder are borne by the child's family,[105] and inclusion is in line with recent publications in this field.[96]
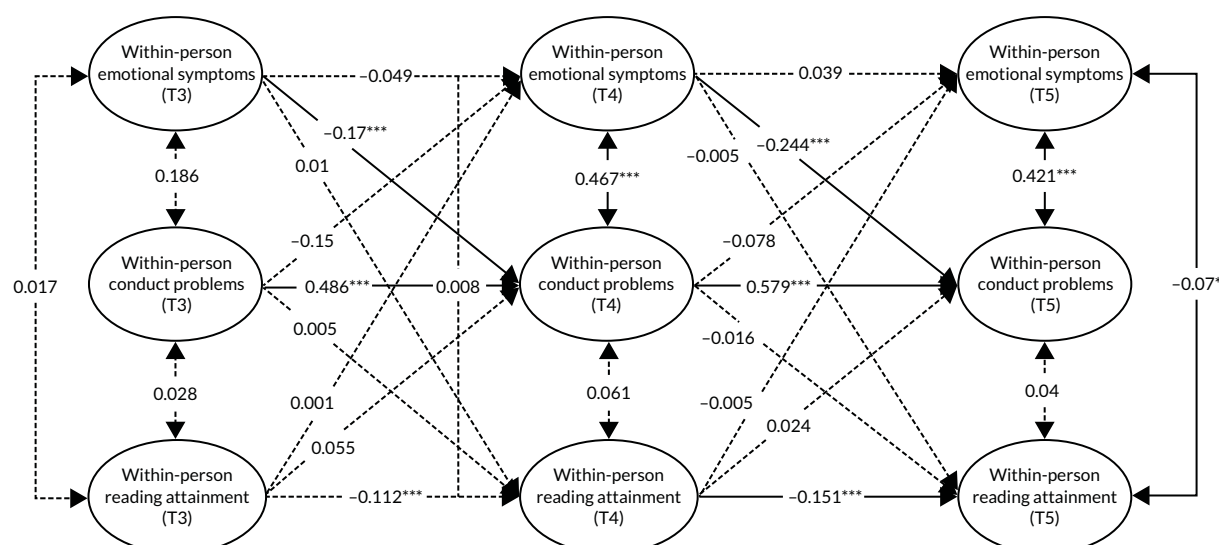
FIGURE 3 Within-person lagged and cross-lagged effects for emotional symptoms, conduct problems and reading attainment (number of participants = 2987), using R. ***$p < 0.001$; standard arrow, statistically significant pathway; dashed arrow, non-significant pathway.

A CCA provides a descriptive analysis of the costs and outcomes of competing alternatives, which is summarised in a CCA balance sheet. The approach allows the decision-maker to draw their own conclusions on the balance of costs and benefits across the alternatives. CCA is the preferred method for assessing the impact of public health interventions[106] and is increasingly used in studies exploring behavioural interventions.[96,107] In this analysis, costs were limited to GBG implementation costs and an estimate of the indirect costs related to exclusion from school. GBG impact is reported based on the primary study outcome (conduct problems; hypothesis 1a) and seven secondary outcomes [psychological well-being, emotional symptoms, peer and social support, school environment, school absence, bullying (i.e. social acceptance) and exclusion from school; hypotheses 1b–h, respectively]. Results are presented as disaggregated costs and outcomes in a cost–consequences table.

Analysis inputs

The Good Behaviour Game cost inputs
The implementation costs of the GBG were provided by Mentor UK in the form of a summary spreadsheet. Additional explanation of the GBG cost components was not provided. Cost components included recruitment and school identification set-up costs in addition to GBG implementation costs. These trial-related costs were excluded from our analysis (*Table 23*).

Indirect cost inputs
Exclusion data for the GBG trial cohort derived from the NPD were provided by ONS. The data were aggregated to the level of intervention/control. The data could not be reported at a more granular level (e.g. school, class, or pupil) because of privacy protection reporting restrictions[87] imposed by ONS that prevent reporting of outcomes for which $n < 10$. Given the ages of the pupils in the trial cohort, absence due to exclusion was assumed to require parent or carer supervision. The cost of time was proxied to the median wage based on historical earnings data.[108] Three cost years (i.e. 2015, 2016, 2017) were included to cover the two school years. Average salaries were not used as salary data are non-parametrically distributed and heavily skewed; therefore, the median provides a better measure of central tendency. One exclusion comprised one half-day absence from school. An average cost per session was calculated (*Table 24*).

TABLE 23 GBG implementation costs (data provided by Mentor UK; total of included cost components £430,068)

Reported cost components (£)	Budget year			
	2014/15[a]	2015/16	2016/17	2017/18[b]
Included cost components (related to GBG implementation)				
Head coach	1680	0	30,720	0
7 coaches	0	85,238	127,906	26,578
Travel: head coach	357	0	11,383	0
Travel: coaches	62	0	4990	4739
Travel not yet assigned	0	22,922	0	0
Manuals	0	12,358	2925	0
Materials	0	27,659	11,517	0
Licence	6742	0	0	0
Teacher training	0	21,961	14,908	0
AIR labour and travel	0	7778	7645	0
Excluded cost components (assumed related to set-up or conduct of the trial)				
Team meeting	0	0	2025	89
Travel: project director	980	0	6220	0
Travel: project officer	441	0	6759	0
Staff recruitment	36	1045	0	0
School recruitment	5884	0	395	0
Incentives	0	69,161	7150	6900
Public relations and communications	2696	1458	0	0
Sundries	198	3062	708	21

a Costs incurred in the 2014/15 financial budget are assumed to be costs related to the start-up of the programme.
b Academic years do not run concurrently with financial years, and costs paid from the 2017/18 financial budget would probably have been incurred in the 2016–17 academic year.

Note
Data were provided by Mentor UK in 2019; the data were provided without additional explanation of time incurred versus time billed. Data in the table represent costs logged as 'actual' across each budget year.

TABLE 24 Estimate of indirect costs related to exclusion, to proxy parent/carer income loss

Inputs	Value	Source
Median weekly salary		
2015	£528	ONS Annual survey of earnings 2015[109]
2016	£539	ONS Annual survey of earnings 2016[110]
2017	£550	ONS Annual survey of earnings 2017[111]
Average	£539	Calculated
Number of school sessions per week	10	Two sessions per school day
Cost per excluded session	£53.90	Calculated

The Good Behaviour Game outcome data
Other than the data for exclusion from school, the GBG outcome data (hypotheses 1a–g) were taken from the main analyses and are reported directly in the cost–consequences balance sheet (*Table 25*).

Analysis outputs

The Good Behaviour Game intervention costs
The GBG intervention costs were estimated by school, pupil, and cost per pupil per year (*Table 26*).

TABLE 25 Cost–consequences balance sheet: per pupil

Parameters	GBG	Usual practice	Difference	*p*-value
Summary of costs				
Per school				
GBG	£11,318	£0	£11,318	n/a
Exclusion from school	£228.37	£234.95	−£6.58	n/a
Per pupil				
GBG	£275.68	£0	£275.68	n/a
Exclusion from school	£5.59	£6.07	−£0.48	n/a
Summary of consequences				
Conduct problems (hypothesis 1a), % (*n*)	13.05 (157)	12.6 (165)	−0.039 (0.323)	0.903
Psychological well-being (hypothesis 1b), mean (SD)	48.617 (10.574)	49.209 (10.098)	−0.251 (0.354)	0.477
Emotional symptoms (hypothesis 1c), % (*n*)	10.22 (123)	10.99 (144)	−0.265 (0.318)	0.405
Peer and social support (hypothesis 1d), mean (SD)	51.721 (11.833)	51.699 (11.236)	−0.016 (0.356)	0.965
School environment (hypothesis 1e), mean (SD)	53.208 (10.902)	53.312 (10.657)	0.019 (0.319)	0.952
School absence (hypothesis 1f), mean % (SD)	95.69 (5.16)	95.73 (5.27)	−0.065 (0.046)	0.154
Bullying (i.e. social acceptance; hypothesis 1g), mean (SD)	45.932 (11.755)	46.009 (12.183)	0.191 (0.393)	0.627
Exclusion from school (hypothesis 1h), mean	0.10	0.11	−0.01	0.866

n/a, not applicable.

TABLE 26 The GBG implementation costs (£)

Total	Per school (*n* = 38)	Per pupil (*n* = 1560)	Per pupil per year
430,068	11,318	275.68	137.84

Based on the tally of costs in *Table 23*. Per pupil per year cost was based on 2 years of intervention exposure. Mentor UK had planned a real-world roll-out of the GBG, intended to include a 'grouped cost', which might have resulted in lower cost across each of the calculated metrics.[1]

Exclusion-related indirect costs

The indirect costs of exclusion were calculated based on an estimated cost per exclusion of £53.90 (see *Table 24*). The comparative analysis (*Table 27*) was consistent with earlier analyses of the impact of the GBG on exclusion from school, revealing no substantive difference by trial arm. In accordance with this, differences in exclusion-related costs are minimal.

Cost–consequences balance sheet

The cost–consequences balance sheet summarises the cost and outcomes over the intervention period. The GBG study resulted in estimated implementation costs of £275.68 per child, with no attendant difference found in primary or secondary outcomes (hypotheses 1a–g) and no difference in exclusion cost. Consequently, we found no evidence to support hypothesis 6.

TABLE 27 Indirect costs of exclusion in the GBG trial

Intervention	Pupils (*n*)[a]	Sessions excluded		Exclusion-related costs (£)		
		Total (*n*)	Mean (SD)/mean (95% CI; *p*-value)	Total	Per school	Per pupil
GBG	1552	161	0.1037 (SD: 1.3675)	8677.90	228.37	5.59
Usual practice	1509	170	0.1127 (SD: 1.5444)	9163.00	234.95	6.07
Difference	–	–9	–0.0089 (–0.112 to 0.094; 0.866)	–485.1	–6.58	–0.48

a NPD could not match 23 pupils.

Chapter 4 Discussion

Principal findings

The primary aim of this study was to examine the impact of the GBG on a range of outcomes for children in primary schools in England. Our objectives were to determine (1) the impact of the GBG on health- and education-related outcomes for children; (2) the impact of the GBG on these outcomes for boys at risk of developing conduct problems; (3) the extent to which the effects of the GBG vary by intervention compliance (i.e. dosage); (4) whether or not the effects of the GBG are sustained (or emerge) over time; (5) the temporal association between mental health and academic attainment; and (6) the costs and consequences of the GBG.

In relation to objective 1, no evidence of the impact of the GBG was found in our ITT analyses. Similarly, for objective 2, our subgroup moderator analyses revealed no impact of the GBG on any outcomes for boys at risk of developing conduct problems, with the exception of a significant negative effect on bullying. With regard to objective 3, there was minimal and conflicting evidence regarding the effects of intervention compliance. Thus, although moderate and high compliance produced significant reductions in school absence, it also led to significant reductions in psychological well-being. Analysis of data pertaining to objective 4 revealed no evidence of the emergence of intervention effects at the 12- or 24-month follow-ups on any outcomes, with the exception of a negative effect on peer and social support. For objective 5, after estimating within- and between-individual effects, we found no temporal associations between children's mental health and their academic attainment. Last, our CCA for objective 6 indicated that the GBG does not provide value for money.

Hypothesis 1: main intervention effects (intention to treat)
The lack of any intervention effect across eight different health- and education-related outcomes (hypothesis 1) appears to provide a robust indictment of the GBG, at least from the perspective of the most bias-free analytical framework: ITT. Our findings align with those of the EEF-funded trial on which the current study was built, which found no impact of the intervention on children's reading attainment, disruptive behaviour, concentration problems or prosocial behaviour.[1] As this is the first trial of the GBG in England, cultural incompatibility cannot be ruled out. Aspects of the IPE reported in the EEF-funded trial support this assertion; for example, many teachers reported struggling with certain mandated intervention procedures, most notably not being able to directly interact or intervene with pupils during gameplay.[1]

However, the initial pilot in Oxfordshire concluded that, on the whole, the GBG was acceptable and feasible in the English school context,[52] and so an apparent lack of cultural transferability cannot be the only explanation. Furthermore, there are also parallels with two recent US-based trials,[18,37] which also found no main effect of the GBG on a range of outcomes (note that although the Ialongo *et al.*[37] trial did report significant intervention effects, this was for the combined 'PATHS to PAX' condition only; in the 'GBG only' condition, no main effects were identified).

These two recent trials notwithstanding, the majority of studies outlined in *Table 1* that undertook true ITT analyses (recalling that three trials[17,23,33] reported subgroup moderator effects only) identified at least one significant intervention effect (albeit typically with small or moderate ESs). What then are we to make of this apparent divergence from a clear trend in the evidence base for the GBG? The discrepancy appears unrelated to the nature of the outcomes assessed, given that these other trials have assessed similar domains to those reported here, including sometimes using the same instruments (e.g. the trial by Jiang *et al.*[19] used the teacher-informant report version of the SDQ from which our primary outcome of conduct problems and secondary outcome of emotional symptoms are drawn).

Similarly, it does not appear to be an artefact of the specific version of the GBG used here, as the two Dutch trials also implemented the AIR model and reported positive effects.[31,32] Despite the fact that it is often trialled in combination with other interventions,[17,29,35] the fact that the GBG was implemented in isolation in the current study cannot account for our null findings, given the multiple cases of positive effects where this is also the case.[19,21,31,32,34,36] Furthermore, at 2 years, the overall period of implementation was clearly sufficient, given that effects of the GBG have been found after as little as 10–12 weeks.[36,38]

One immediate possibility is that, as noted in *Chapter 1*, children's behaviour is typically very good in most schools, with very few displaying the symptoms of conduct or other problems at the outset of any given trial. The lack of main intervention effect may, therefore, simply reflect a low base rate of conduct problems in our sample ($\approx$ 16% with scores in the borderline or abnormal range at baseline; and $\approx$ 60% with scores of 0 at baseline). Unfortunately, there is insufficient information provided in the Jiang *et al.*[19] study to enable a direct comparison in terms of base rate and scores for conduct problems, and so this explanation cannot be ruled out. Beyond this, there are three substantive explanations (suboptimal implementation, insufficient programme differentiation and delayed effects) that we focus on in our discussion of findings relating to hypothesis 3 [see *Hypothesis 3: implementation effects (dosage)*] and hypothesis 4 [see *Hypothesis 4: maintenance/sleeper effects (12- and 24-month post-intervention follow-ups)*].

Hypothesis 2: subgroup effects (boys at risk of developing conduct problems)

In relation to hypothesis 2, the lack of evidence for subgroup moderator effects relating to our subsample of at-risk boys across the range of trial outcomes conflicts with the trial evidence on which our hypothesis was based.[23,25] However, that study was set in a single, socioeconomically deprived American city (i.e. Baltimore) at a time when it was beset by crime, substance abuse, antisocial behaviour and myriad other social problems. By contrast, the current trial spanned 23 local authorities across 3 regions of England. Although this encompassed a very diverse range of settings, few, if any, could be argued to parallel the challenges faced by children and families in 1980s inner-city Baltimore. Thus, the 'at-risk boys' subgroups in the two trials were probably qualitatively different in respect of presenting patterns of aggressive and antisocial behaviour and the factors underpinning them. In other words, our subsample of at-risk boys may not have been sufficiently at risk to reap the benefits observed in the Baltimore trial.

In the context of the current study, it is noteworthy that teachers' views on who benefited more (or less) from the GBG, solicited and reported in the EEF trial IPE,[1] were mixed. Although some provided accounts indicating that the intervention had been more beneficial for boys whose behaviour was a cause for concern (consistent with hypothesis 2), others instead highlighted differential gains for children with distinct needs (e.g. those with autism), and several reported that they felt that any benefits were not specific to any particular children in their classroom. These findings align broadly with the subgroup moderator analyses reported here, in that they indicate a lack of clear and distinct benefit for boys at risk of developing conduct problems across classrooms in the intervention arm of the trial, at least to the extent that implementing teachers would consistently identify this in their accounts of the impact of the intervention.

Nonetheless, our findings have important implications for the utility of the GBG as an efficacious means of preventing the maintenance or escalation of conduct problems from childhood to early adolescence, a current priority for both the Government[8] and NICE.[9] On the basis of the evidence accrued in our hypothesis 1 and hypothesis 2 analyses, this universal intervention should not be recommended as an efficacious approach, either to prevent the development of conduct problems (hypothesis 1) or to address existing conduct problems (hypothesis 2). Indeed, the only noteworthy finding in our subgroup moderator analyses was a potential negative intervention effect: at-risk boys in GBG schools reported significantly increased experiences of bullying. Although unexpected, there is a plausible explanation for this potential iatrogenic effect of the GBG. As noted in our description of

the intervention (see *Chapter 2, Methods*), the GBG uses an interdependent group contingency model; in other words, the provision of rewards/reinforcement is contingent on all members of a given group abiding by the rules. This approach may inadvertently evoke negative peer pressure towards those who frequently break the rules (e.g. our at-risk boys subsample) and thereby reduce the likelihood of their group winning the game.[112] This aspect is a hypothesis for a clearly described and testable causal mechanism that future research could explore (including whether or not the effect is sustained in the short to medium term).

In the interim, this unanticipated negative intervention effect suggests that caution is warranted among those who might consider using the GBG. Although the associated ES is very small when judged by conventional standards,[42] it is important to consider what it means in less abstract terms.

First, by transforming our standardised ES (Hedges' *g*) to Cohen's U^3 [using the formula $U^3 = \Phi(\delta)$, in which Φ is the cumulative distribution function of the standard normal distribution, and δ is the population parameter of Cohen's *d*/Hedges' *g*], we can conclude that the subgroup effect of the GBG approximates to a 5-percentile point increase in bullying. Put another way, the observed ES of –0.125 means that 55% of at-risk boys in GBG schools will have a bullying score that is below the mean of those in usual-practice schools (remembering, of course, that a lower score reflects more frequent bullying in the measure used, and that, with no intervention effect whatsoever, we would naturally expect 50% to have a lower score).

Second, by referring back to *Table 4*, we can see that the negative intervention effect identified reflects an approximate decrease of just 1.5 points (SD of 12 multiplied by standardised ES of –0.125) in the possible scoring range of 0–100. For reference, a qualitative shift in the average reported frequency of bullying (e.g. from 'seldom' to 'quite often', or 'quite often' to 'very often' at the item level) would be between 9 and 19 points on the Rasch normalised scores used in the Kidscreen survey.[70]

Last, the average scores for both the trial sample overall and the at-risk boys subgroup at T3 equated to being bullied somewhere between 'never' and 'seldom'. Given all of the above, we conclude that, although the negative intervention effect observed is clearly real and potentially uncomfortable for the affected participants, its magnitude is unlikely to reflect significant harm. Despite this, we are mindful of the fact that this effect, however small, was observed among a particularly vulnerable subgroup of the population who already probably experience low levels of social acceptance. As such, we repeat our note of caution about the future use of the GBG based on the results reported here, unless adaptations can be made to intervention procedures that prevent the negative peer pressure hypothesised above from being the source of this unexpected finding.

This finding aside, the lack of differential gains among our at-risk subgroup provides evidence in support of a more general concern about universal interventions such as the GBG, which is that they lack the intensity to produce meaningful change for children already at risk for psychopathology.[113] This may particularly be the case in the current trial, in which the average dosage was markedly lower than both that recommended by the developer[53] and the dosage that has been reported in other trials (albeit using teacher self-report in most cases).[18,19,36,37] One might tentatively predict that optimal intervention exposure matters most for those who are already at risk. The social adaptation process through which the GBG is believed to impact on behaviour is cumulative in nature, and those at risk are arguably in greater need of the increased opportunities for reinforcement, consolidation and generalisation of learning associated with increased levels of exposure, as this will mitigate the lack of adaptive socialisation in other developmental contexts. This is a point to which we return when discussing directions for future research at the end of *Chapter 5*.

Hypothesis 3: implementation effects (dosage)

As noted above, intervention compliance in the trial, specifically dosage, was suboptimal. The GBG was played for an average of 1066 minutes in total in classrooms in the intervention arm of the trial during

the 2 years of implementation (i.e. approximately 530 minutes in the first year and 524 in the second year). Teachers played the game between once and twice a week, for an average of approximately 25 minutes per week in total. Furthermore, nine schools formally ceased implementation prior to the end of this main trial period. These data contrast sharply with dosage estimates in other trials. For example, teachers in Ialongo et al.'s recent US-based trial[37] reported playing the game for between 1432 (GBG only) and 1583 (combined GBG and PATHS) minutes on average in a single year of implementation. Furthermore, most of the teachers in Streimann et al.'s Estonian GBG trial[21] reported playing the GBG every day; in Tolan et al.'s study,[18] they described playing the GBG more than twice a day on average, for a total of nearly 80 minutes a week. A key caveat with these higher dosage estimates is that they rely on teachers' self-reported implementation behaviour, which is known to be subject to positive bias and impression management.[40] Indeed, a key reason that we developed the bespoke online 'scoreboard' tool was to provide more robust, objective and accurate dosage estimates.[114] However, even if one assumes some level of 'uplift' in the teacher-reported estimates, the total duration reported here is still likely to be significantly lower than the actual level of dosage in the other trials.

Given the above, implementation failure provides a potential explanation for the null results in hypothesis 1. However, our CACE analyses do not appear to support this proposition, as we found no moderate- or high-compliance effects for the overwhelming majority of outcomes (the exceptions being a negative effect on well-being and a positive effect on school absence). In other words, the intervention was found to be ineffective even after robustly accounting for implementation variability. These findings align with those of Bradshaw et al.,[47] whose ITT and CACE analyses focusing on an at-risk subsample also returned null results for the GBG when implemented in isolation (as opposed to in combination with PATHS, for which there were significant intervention effects in the ITT analysis that grew in magnitude once compliance was taken into account). However, before ruling out implementation failure, we must consider the possibility that a 'minimum effective dose'[115] was not reached in even our high-compliance classrooms. In these settings, children were exposed to the GBG for > 1348 minutes across the 2-year implementation period. This is still significantly lower than the average cumulative intervention intensity achieved in a much shorter period in the above trials.[18,21,37] Thus, in the absence of a minimum effective dose, generalisation of learning (and consequent changes in behaviour and other outcomes) beyond the immediate context of the game itself may not follow.[116]

In accordance with this, it is worth briefly considering the findings from the EEF trial IPE,[1] which highlighted a range of factors affecting implementation. Notably, these included pupil needs (e.g. teachers were more likely to play the game frequently if they believed that there were children in their classroom for whom it would be beneficial), teacher attitudes (e.g. teachers were less likely to play the game frequently if they felt that it was difficult to integrate with their lesson plans), and competing priorities (e.g. teachers were less likely to play the game frequently during busy periods of the school year in which there were scheduled school events and/or assessments). The first of these qualitative findings (pupil needs) is supported by the CACE models reported in the current study, in as much as individual- and school-level characteristics indicative of need (e.g. conduct problems, prosocial behaviour) were found to be predictors of compliance. Taken together, the findings indicate that, in some contexts, such as those where pupil need is considered to be low, achieving a minimum effective dose may not be a realistic objective; in other contexts, where need is high but attitudinal or logistic barriers are present, there are implications for the training and support model used in the GBG (e.g. additional input to 'win hearts and minds' and problem solve implementation challenges).

Despite the above, we remain resolute in our conclusions about the implications for the GBG as a means of preventing the maintenance or escalation of conduct problems and other maladaptive outcomes in the English school context. This is because the dosage reported in the current trial is likely a 'best-case scenario' for implementation here, as it was achieved in an efficacy trial context in which initial training and ongoing coaching support for teachers, subsidised intervention costs for schools, additional provision for data monitoring from our research team, and developer (AIR) support for the delivery team, Mentor UK, were all available.[1] In other words, although we may have seen more

evidence of meaningful intervention effects with significantly higher levels of implementation than those that were observed here, it is very unlikely that such levels would ever realistically be achieved were the GBG to be implemented at scale in England, in which case such a comprehensive implementation support system would be absent.

Setting aside insufficient implementation, a further potential explanation for our overall pattern of findings relates to programme differentiation. Recall that our survey of teachers' behaviour management strategies and approaches revealed that those in the control arm of the trial were enacting practices that mirrored some of the core components of the GBG (e.g. classroom rules, team membership, monitoring behaviour and positive reinforcement). Given this, it is possible that the null results observed were due to the fact that the intervention was insufficiently differentiated from the usual practice of schools. This proposition appears to be supported by a general trend in which null findings are more commonplace in recent GBG trials (see *Table 1*). This may be the result of the behaviour management practices that are central to the intervention reducing in novelty and becoming more endemic over time (although we note that this explanation is confounded by the increased rigour of trials over time; more frequent null results in recent years may also be the result of more robust testing of the intervention). It may also explain why the GBG is so often delivered in combination with other interventions.[17,18,35,37,38] That is, in isolation, the GBG is more likely to be ineffective because of limited programme differentiation, but when implemented in conjunction with other preventative interventions, multiplicative effects are observed as a consequence of the interaction of complementary active ingredients.[117]

Hypothesis 4: maintenance/sleeper effects (12- and 24-month post-intervention follow-ups)

Notwithstanding the preceding explanations for the lack of intervention effects, a further possibility is that they simply had not yet emerged at the end of the main implementation period. Preventative effects may take time to emerge, especially when a relatively small proportion of the population has (or is at risk of developing) problems in the first place.[118] Thus, comprehensive evaluation requires outcome assessment to go well beyond the cessation of a given intervention for changes among intervention recipients to consolidate, for small but key changes to snowball and for the members of the control group to exhibit difficulties of the kind that are the focus of prevention efforts.[119] Although longer-term follow-up studies of the GBG are scarce, those that were conducted prior to the current trial each found intervention effects (albeit following improved outcomes in the short-term; hence, they all establish maintenance rather than sleeper effects; see *Table 1*). However, this was not the case here. Indeed, the only notable effect in our post-intervention follow-up analyses was an unexpected one, whereby children in the intervention arm reported significantly lower levels of peer and social support than their peers in the control arm 2 years after the end of the main trial. Thus, although it is technically possible that sleeper effects may yet emerge in our trial cohort, this seems highly unlikely. As a consequence, we feel confident in ruling out the timing of effects as a contributory factor in the results of this study.

Hypothesis 5: the temporal association between mental health and academic attainment

The design of the current study afforded an opportunity to contribute to the growing literature on the temporal associations between health and educational outcomes. Specifically, research in developmental cascades has documented, over varying lags, ranging from 12 months[5] to many years,[56] the apparently reciprocal longitudinal relations between domains of mental health and academic attainment. We found no evidence of this in our analyses pertaining to hypothesis 5. However, it is important to note that most prior work in this area has used cross-lag panel models that are unable to disaggregate between- and within-individual effects. This means that the apparent cross-domain within-individual temporal effects reported in these studies may be erroneous.[86] In the current study, we used the RI-CLPM approach, finding clear trait-like between–individual associations (e.g. children experiencing emotional symptoms and/or conduct problems tend to also have lower reading scores), but no evidence of genuine within-individual temporal associations between either domain of mental health and academic attainment (albeit with relatively poor model fit). Thus, after controlling for shared risk and between-individual

effects, we do not find support for the hypothesis that mental health influences later attainment, or vice versa, at the end of middle childhood. Whether or not our analysis is an outlier remains to be seen; further research that uses RI-CLPM in this space is needed.

Hypothesis 6: costs and consequences of the Good Behaviour Game

Our health economic analysis found no evidence to support the argument that the GBG provides value for money. The total implementation costs of the intervention were calculated at £11,318 per school over the main trial period (a cost of £275.68 per pupil). This intervention cost was the driver in cost differences in the trial, with the only other captured cost, exclusion from school, showing little difference (£0.48 per pupil) over the same time period. Examining the outcomes at deeper levels of granularity (e.g. mapping costs to level of compliance and/or dose) was not possible because of the aggregated nature of the available data.

It should be noted that our reported outcomes are based on a single year group in each intervention school being exposed to the GBG, whereas, in a real-world setting, a roll-out across the whole school might be expected, resulting in a lower cost per pupil. Hence, the main report of the EEF-funded GBG trial estimated that the cost of implementing the GBG within primary schools was similar to that noted here, £11,000 per school over a 3-year period, comprising £4500 start-up costs in the first year and £3000 per annum in the subsequent years, but a per-child implementation cost of £37 that assumed roll-out across the whole school.[1] In addition, it should be noted that Mentor UK's original plan for roll-out of the GBG programme was to include a discount for grouped school uptake, which may have resulted in lower costs; their costs were modelled on a 10-school cohort, and larger cohorts would probably incur lower costs owing to shared resources (e.g. coaches).

Previous studies of the GBG have indicated strong performance and good value for money, with the Washington State Institute for Public Policy (WSIPP) most recently reporting a 60-fold return on an investment, that is a return of $63 for every $1 spent implementing the intervention.[120] However, although targeted parent-level interventions for the prevention of conduct problems have shown good economic return in the UK,[121] in general, strong economic arguments for UK-based classroom-level interventions remain limited. Recent studies have highlighted the challenge of capturing the benefits of such approaches within a standard economic framework,[96,122–124] with the per-child cost estimates of trialled interventions ranging from £8[124] to £153.[123] These findings are not directly comparable with the GBG cost estimates reported here because of the limited granularity of the data.

Owing to the reporting restrictions on exclusions data,[87] it is not possible to elicit, in either arm of the trial, whether these exclusions represented one child being excluded many times or many children being excluded once; this is information that could be important when assessing the impact of the GBG on exclusion from school. Further exploration of findings could help to focus the intervention on schools that are most likely to benefit. It is also possible that there are other measurable costs that were not captured by the study that may be important to the health economic assessment of the GBG (e.g. health and social care utilisation); however, given the lack of available data, the decision on whether or not the intervention should be implemented further is likely to be made using the consequence outcomes (hypotheses 1a–g).

Non-hypothesised findings

In the course of our analyses, some notable findings emerged that did not directly relate to our prespecified trial hypotheses. In keeping with the intention to avoid mining/dredging, these have been compartmentalised from the main findings discussed above and are noted briefly here, with a very clear caveat that they should be considered as exploratory.

The first finding relates to subgroup moderator effects. Modelling our hypothesised subgroup effect (differential effects for at-risk male pupils) required us to fit a three-way interaction term (e.g. GBG*male*at risk); doing so meant that we also needed to fit three possible two-way interaction

terms involving the trial group, sex and risk status variables (e.g. GBG*male, GBG*at risk, and male*at risk). In the course of doing so, we observed significant interactions between trial group and risk status that denote reductions in bullying and absence among at-risk students in GBG schools; we also observed a significant interaction between trial group and sex that was indicative of an increase in absence among male pupils in GBG schools (see *Tables 10* and *11*).

The second finding relates to within-individual temporal associations between emotional symptoms and conduct problems. As above, these were not specified in the study hypothesis and are an artefact of the modelling process. Here, we found a consistent pattern in which emotional symptoms were significantly inversely related to later conduct problems. From T3 to T4, an increase of one SD in emotional symptoms was associated with just less than one-fifth of a SD decrease in later conduct problems. This effect increased in magnitude from T4 to T5, with a one SD increase in emotional symptoms being associated with approximately one-quarter of a SD decrease in later conduct problems. This finding supports the proposition that the experience of emotional symptoms may serve a protective function in relation to later conduct problems by interrupting the trajectories from risk to disruptive behaviours, perhaps owing to the increased self-isolation and withdrawal associated with emotional symptoms.[125]

Strengths and limitations

The current study has numerous strengths that increase the security of our findings. First, a cluster-randomised design with appropriate analyses that took account of the hierarchical and clustered nature of the data set was used. Second, the trial was very large and well powered; furthermore, the 77 trial schools spanned 23 local authorities across 3 regions of England, thereby providing a much greater diversity of settings than most other studies of the GBG. Third, attrition was within acceptable limits, being 0% at the school level and 18.5% at the child level at the point of our main ITT analyses, and missing data were accounted for using FIML, eliminating the bias associated with complete-case analysis. Fourth, the use of a randomised design (in which allocation was determined by an independent trials unit) meant that, in expectation, we would be free from confounders. Fifth, the use of a cluster-randomised design and the proprietary nature of the GBG minimised the possibility of contamination effects. Sixth, our assessment of primary and secondary outcomes used multiple methods (e.g. surveys, standardised tests) and informants (e.g. children, teachers), which was consistent with recommended practices. Last, our design enabled a very thorough and comprehensive assessment of the GBG, including the robust examination of three key effect modifiers: subgroups, implementation and timing of follow-up.

Nonetheless, there are also a number of limitations that need to be considered. First, children and teachers completing outcome measures were not blinded to trial allocation, which potentially introduced bias (although this seems unlikely given the null findings). Furthermore, given the lack of blinding, we cannot rule out compensatory rivalry as a partial explanation for our findings (e.g. increased efforts among usual-practice schools in response to not being allocated to the intervention arm of the trial). Second, independent observation of children's behaviour would perhaps have been preferable to the use of surveys; however, this was not feasible within the resources available for the trial and would also have significantly increased the data burden on participating schools. Third, our analyses were reliant on point-in-time estimates, and it has been argued that these do not provide a fair test of the efficacy of preventative interventions, which are designed to alter children's developmental trajectories.[119] Fourth, intervention compliance was suboptimal, and, although our CACE analyses indicated that outcomes mostly did not vary as a function of dosage, we cannot rule out the possibility that a minimum effective dose was not reached, even in our high-compliance settings. Fifth, our CACE analyses were specifically related to dosage, meaning that other potentially important dimensions, such as procedural fidelity, were neglected (although these appeared to be less variable and much closer to optimal levels than dosage). Sixth, the 'augmented' nature of the trial meant that, for several of our secondary outcomes, a post-test-only design was used rather than the generally preferred pre-test-post-test design that was used for our primary outcome. However, as a counterpoint, we note Gorard's argument[126] that the post-test-only

design is 'generally at least as safe as its alternatives'. Last, as for any study of this kind, attrition occurred and increased over time (e.g. 80% of participants with primary outcome data for both T1 and T3; 76% with data for both T1 and T5). However, appropriate techniques to account for missing data (e.g. FIML) were employed where possible, meaning that the majority of models were based on the full trial sample ($n = 3084$), thereby reducing the bias associated with attrition.

Chapter 5 Conclusions

Main findings

1. No evidence of the impact of the GBG was found in our ITT analyses.
2. Our subgroup moderator analyses similarly revealed no impact of the GBG on any outcomes for boys at risk of developing conduct problems, with the exception of a significant but small increase in bullying.
3. There was minimal and conflicting evidence regarding the effects of intervention compliance. Thus, although moderate and high compliance produced significant reductions in school absence, they also led to significant reductions in psychological well-being.
4. No evidence of the emergence of intervention effects was found at 12- or 24-month follow-up for any outcome, with the exception of a negative effect on peer and social support.
5. After estimating within- and between-individual effects, we found no temporal associations between children's mental health and their academic attainment.
6. Our CCA indicated that the GBG does not provide value for money.

Implications

Developing the evidence base regarding the most effective behaviour management strategies has been set as a research priority by both the Government[8] and NICE.[9] On the basis of the findings reported here, it is not possible to recommend the GBG as an efficacious means through which teachers can manage the behaviour of and improve health- and education-related outcomes for children and young people. Our findings demonstrate that the GBG is not superior to existing practice.

In accordance with this, those seeking to adopt behaviour management strategies for which there is clear evidence of positive intervention effects generated from robust trials conducted in England should consider alternative approaches. For example, the efficacy of the Incredible Years Teacher Classroom Management programme was recently tested in a large RCT in England, reporting short-term improvements in mental health, peer relationships and prosocial behaviour, alongside short- to medium-term improvements in overactivity and disruptive behaviour.[96] Recently published behaviour management guidance for schools,[127] underpinned by a comprehensive review of the research literature,[128] may also be useful.

Recommendations for future research

In this final section, we consider outstanding research questions and gaps arising from the current study, alongside potential methodological/analytical developments, which may lead to greater evidential insights into how to effectively manage behaviour in the classroom and, in particular, prevent the maintenance or escalation of conduct problems from childhood to early adolescence.

In terms of outstanding research questions and gaps arising from the current study, it might be useful for future research to more formally explore the issue of programme differentiation when complex psychosocial and behavioural interventions are imported into different settings. In the > 50 years since the first report of the GBG was published, many of its constituent components have become commonplace behaviour management techniques and, as noted in *Chapter 4*, our survey of teachers' behaviour management strategies and approaches in the current trial revealed that those in the control arm were enacting practices relating to classroom rules, team membership, monitoring

behaviour and positive reinforcement. Future research should therefore examine whether or not the magnitude of intervention effects varies by level of programme differentiation. One might, for example, predict larger effects in 'high-differentiation' settings, in which the constituent components of the GBG are novel, than in 'low-differentiation' settings, in which they are less distinct from existing practice.

Building on the above, and given the relative frequency with which the GBG has been trialled in combination with other preventative interventions (e.g. curriculum enhancements, social and emotional learning programmes, peer tutoring; see *Table 1*) in US-based trials, future research in England could examine whether or not an integrated model (e.g. the GBG plus a complementary intervention) could produce favourable health- and education-related outcomes. As noted earlier, this is plausible if we assume that, in isolation, the GBG is more likely to be ineffective because of limited programme differentiation, but, when implemented in conjunction with other preventative interventions, multiplicative effects are observed as a consequence of the interaction of complementary active ingredients.[117] This supposition, in turn, leads to a further objective for future research to address: identifying these active ingredients. This could involve new primary research (e.g. factorial trials using the multiphase optimisation strategy framework, in which different combinations of intervention components are varied across trial arms) and/or aggregative reviews of existing evidence (e.g. meta-analyses in which the interventions are coded for the presence/absence of different intervention components, which in turn are used as moderators in metaregression models).

Turning now to methodological/analytical developments, recall that, in line with existing literature, treatment effect modifiers (e.g. implementation variability, differential gains among subgroups, timing of effects) were assessed independently in the current trial. Future research should begin to examine their integration. This might involve, for example, extensions of CACE models to include subgroup moderator analyses, or medium- and long-term follow-up data points. In the case of the subgroup moderator analyses, such models would allow researchers to examine who benefits more (or less) from a higher dosage. At the time of writing, members of our team have recently published a CACE analysis using data from the EEF trial that revealed significant effects of the GBG on teacher-rated disruptive behaviour, with differential gains evident among children at varying levels of cumulative risk exposure in the context of moderate and high compliance.[129] In the case of the medium- and long-term follow-up data points, these would allow researchers to identify sleeper effects that are conditional on exceeding a particular dosage threshold. As noted in *Chapter 1*, such effects have already been identified using data from this study: Ashworth *et al.*[48] revealed sleeper effects of the GBG on academic attainment at 12-month post-intervention follow-up among compliers.

A noteworthy feature of the current trial (and of much of the literature) is a reliance on point-in-time estimates (i.e. the impact of the GBG on a given outcome at a particular point in time) that do not analyse the developmental process of growth. As a key purpose of universal interventions is to alter developmental trajectories, it is important that this is reflected in the analytical techniques adopted by researchers.[119] Future trials of preventative school-based interventions might therefore use growth curve models more frequently. Members of our team have begun to apply such models to data from the current study, with promising results (e.g. a noteworthy impact of the GBG on developmental trajectories of concentration problems and prosocial behaviour from T1 to T5).[130] Following the above theme of integration, one could also envisage the utility of growth curve models that incorporate compliance information to examine whether or not developmental trajectories of behavioural and other outcomes vary by intervention dosage.

Last, the lack of reciprocal longitudinal relations between domains of mental health and academic attainment in our analysis pertaining to hypothesis 5 prompts further research that can determine whether our findings are outliers or it is simply the case that previous studies have inadvertently reported such associations because their analyses were unable to discriminate within- and between-individual effects.[86] Therefore, further research that makes use of the RI-CLPM framework is needed, including the re-examination of already published analyses and/or new 'side-by-side' comparisons of

RI-CLPM and traditional CLPM. Where such work has already taken place in relation to other theorised longitudinal associations between psychological and attainment outcomes, there is evidence that traditional CLPM overestimates causal paths. For example, Burns *et al.*[131] found that reciprocal longitudinal associations between self-concept and academic attainment established in CLPM were partially or fully attenuated in RI-CLPM. Given that theorised longitudinal relations between domains of mental health and academic attainment have become an important element of the discourse in relation to the importance of mental health promotion in schools, clarification is urgently required. However, we would also caution that, even if future research using RI-CLPM affirms the findings reported here, this does not undermine the argument for school mental health promotion as important in and of itself. In other words, being longitudinally related to academic attainment is not a prerequisite for mental health to be a fundamental part of schools' remit and responsibilities.

Acknowledgements

This research was made possible first and foremost by the 3084 children who participated in the study; we are extremely grateful to them. We would also like to thank their parents, teachers and schools. The TSC, led by Professor Tamsin Ford (University of Cambridge, Cambridge, UK), AIR and Mentor UK all played important roles for which we express our gratitude. Last, we would like to thank the reviewers of this report for their feedback during the drafting process.

Contributions of authors

Publications

Ashworth E, Panayiotou M, Humphrey N, Hennessey A. Game on – complier average causal effect estimation reveals sleeper effects on children's academic attainment in a randomized trial of the Good Behavior Game. *Prev Sci* 2020;**21**:222–33.

Ashworth E, Humphrey N. Game over? No main or subgroup effects of the Good Behavior Game in a randomized trial in English primary schools. *J Res Educ Eff* 2020;**13**:298–321.

Ashworth E, Humphrey N, Lendrum A, Hennessey A. Beyond 'what works': a mixed-methods study of intervention effect modifiers in the Good Behavior Game. *Psychol Sch* 2019;**57**:222–46.

Humphrey N, Hennessey A, Ashworth E, Frearson K, Black L, Petersen K, *et al. Good Behaviour Game: Evaluation Report.* London: Education Endowment Foundation; 2018.

Humphrey N, Panayiotou M, Hennessey A, Ashworth E. Treatment effect modifiers in a randomized trial of the Good Behavior Game during middle childhood. *J Consult Clin Psychol* 2021;**89**:668–81.

Troncoso P, Humphrey N. Playing the long game: a multivariate multilevel non-linear growth curve model of long-term effects in a randomized trial of the Good Behaviour Game. *J Sch Psychol* 2021;**88**:66–84.

Office for National Statistics disclaimer

This work was produced using statistical data from the Office for National Statistics (ONS). The use of the ONS statistical data in this work does not imply the endorsement of the ONS in relation to the interpretation or analysis of the statistical data. This work uses research datasets which may not exactly reproduce National Statistics aggregates.

Data-sharing statement

This report makes use of demographic and educational data extracted from the National Pupil Database. The authors are unable to make the National Pupil Database data available as they are restricted by their owners (the UK Government). Anonymised data generated through the trial itself (e.g. implementation and outcome data) are available upon request from the corresponding author.

References

1. Humphrey N, Hennessey A, Ashworth E, Frearson K, Black L, Petersen K, *et al. Good Behaviour Game: Evaluation Report and Executive Summary*. London: Education Endowment Foundation; 2018.

2. Department for Education. *Pupil Behaviour in Schools in England*. London: Department for Education; 2012.

3. Office for Standards in Education. *Below the Radar: Low-Level Disruption in the Country's Classrooms*. London: Office for Standards in Education; 2014.

4. NHS Digital. *Mental Health of Children and Young People in England, 2017*. London: NHS Digital; 2018.

5. Panayiotou M, Humphrey N. Mental health difficulties and academic attainment: evidence for gender-specific developmental cascades in middle childhood. *Dev Psychopathol* 2018;**30**:523–38. https://doi.org/10.1017/S095457941700102X

6. Knapp M, King D, Healey A, Thomas C. Economic outcomes in adulthood and their associations with antisocial conduct, attention deficit and anxiety problems in childhood. *J Ment Health Policy Econ* 2011;**14**:137–47.

7. D'Amico F, Knapp M, Beecham J, Sandberg S, Taylor E, Sayal K. Use of services and associated costs for young adults with childhood hyperactivity/conduct problems: 20-year follow-up. *Br J Psychiatry* 2014;**204**:441–7. https://doi.org/10.1192/bjp.bp.113.131367

8. Department for Education. *School Behaviour and Attendance: Research Priorities and Questions*. London: Department for Education; 2014.

9. National Institute for Health and Care Excellence (NICE). *Antisocial Behaviour and Conduct Disorders in Children and Young People: Recognition, Intervention and Management*. London: NICE; 2013.

10. Korpershoek H, Harms T, de Boer H, van Kuijk M, Doolaard S. A meta-analysis of the effects of classroom management strategies and classroom management programs on students' academic, behavioral, emotional, and motivational outcomes. *Rev Educ Res* 2016;**86**:1–38. https://doi.org/10.3102/0034654315626799

11. Tingstrom DH, Sterling-Turner HE, Wilczynski SM. The good behavior game: 1969–2002. *Behav Modif* 2006;**30**:225–53. https://doi.org/10.1177/0145445503261165

12. Donaldson JM, Wiskow KM. The Good Behaviour Game. In Teasdale B, Bradley MS, editors. *Preventing Crime and Violence*. Switzerland: Springer International Publishing; 2017. pp. 229–41. https://doi.org/10.1007/978-3-319-44124-5_20

13. Skinner BF. The operational analysis of psychological terms. *Psychol Rev* 1945;**52**:270–7. https://doi.org/10.1037/h0062535

14. Bandura A. *Social Foundations of Thought and Action: A Social Cognitive Theory*. Hoboken, NJ: Prentice Hall; 1986.

15. Kellam SG, Mackenzie AC, Brown CH, Poduska JM, Wang W, Petras H, Wilcox HC. The Good Behavior Game and the future of prevention and treatment. *Addict Sci Clin Pract* 2011;**6**:73–84.

16. Weis R, Osborne KJ, Dean EL. Effectiveness of a universal, interdependent group contingency program on children's academic achievement: a countywide evaluation. *J Appl Sch Psychol* 2015;**31**:199–218. https://doi.org/10.1080/15377903.2015.1025322

17. Ialongo NS, Werthamer L, Kellam SG, Brown CH, Wang S, Lin Y. Proximal impact of two first-grade preventive interventions on the early risk behaviors for later substance abuse, depression, and antisocial behavior. *Am J Community Psychol* 1999;**27**:599–641. https://doi.org/10.1023/A:1022137920532

18. Tolan P, Elreda LM, Bradshaw CP, Downer JT, Ialongo N. Randomized trial testing the integration of the Good Behavior Game and MyTeachingPartner™: the moderating role of distress among new teachers on student outcomes. *J Sch Psychol* 2020;**78**:75–95. https://doi.org/10.1016/j.jsp.2019.12.002

19. Jiang D, Santos R, Josephson W, Mayer T, Boyd L. A comparison of variable- and person-oriented approaches in evaluating a universal preventive intervention. *Prev Sci* 2018;**19**:738–47. https://doi.org/10.1007/s11121-018-0881-x

20. O'Keeffe J, Thurston A, Kee F, O'Hare L, Lloyd K. Protocol: a feasibility study and a pilot cluster randomised controlled trial of the PAX 'Good Behaviour Game' in disadvantaged schools. *Int J Educ Res* 2017;**86**:78–86. https://doi.org/10.1016/j.ijer.2017.08.003

21. Streimann K, Selart A, Trummal A. Effectiveness of a universal, classroom-based preventive intervention (PAX GBG) in Estonia: a cluster-randomized controlled trial. *Prev Sci* 2020;**21**:234–44. https://doi.org/10.1007/s11121-019-01050-0

22. Smith EP, Osgood DW, Oh Y, Caldwell LC. Promoting afterschool quality and positive youth development: cluster randomized trial of the Pax Good Behavior Game. *Prev Sci* 2018;**19**:159–73. https://doi.org/10.1007/s11121-017-0820-2

23. Dolan LJ, Kellam SG, Brown CH, Werthamer-Larsson L, Rebok GW, Mayer LS, *et al.* The short-term impact of two classroom-based preventive interventions on aggressive and shy behaviors and poor achievement. *J Appl Dev Psychol* 1993;**14**:317–45. https://doi.org/10.1016/0193-3973(93)90013-L

24. Kellam SG, Rebok GW, Ialongo N, Mayer LS. The course and malleability of aggressive behavior from early first grade into middle school: results of a developmental epidemiologically-based preventive trial. *J Child Psychol Psychiatry* 1994;**35**:259–81. https://doi.org/10.1111/j.1469-7610.1994.tb01161.x

25. Kellam SG, Ling X, Merisca R, Brown CH, Ialongo N. The effect of the level of aggression in the first grade classroom on the course and malleability of aggressive behavior into middle school. *Dev Psychopathol* 1998;**10**:165–85. https://doi.org/10.1017/s0954579498001564

26. Kellam SG, Brown CH, Poduska JM, Ialongo NS, Wang W, Toyinbo P, *et al.* Effects of a universal classroom behavior management program in first and second grades on young adult behavioral, psychiatric, and social outcomes. *Drug Alcohol Depend* 2008;**95**(Suppl. 1):5–28. https://doi.org/10.1016/j.drugalcdep.2008.01.004

27. Wilcox HC, Kellam SG, Brown CH, Poduska JM, Ialongo NS, Wang W, Anthony JC. The impact of two universal randomized first- and second-grade classroom interventions on young adult suicide ideation and attempts. *Drug Alcohol Depend* 2008;**95**(Suppl. 1):60–73. https://doi.org/10.1016/j.drugalcdep.2008.01.005

28. Ialongo N, Poduska J, Werthamer L, Kellam S. The distal impact of two first-grade preventive interventions on conduct problems and disorder in early adolescence. *J Emot Behav Disord* 2001;**9**:146–60. https://doi.org/10.1177/106342660100900301

29. Reid JB, Eddy JM, Fetrow RA, Stoolmiller M. Description and immediate impacts of a preventive intervention for conduct problems. *Am J Community Psychol* 1999;**27**:483–517. https://doi.org/10.1023/A:1022181111368

30. Eddy JM, Reid JB, Stoolmiller M, Fetrow RA. Outcomes during middle school for an elementary school-based preventive intervention for conduct problems: follow-up results from a randomized trial. *Behav Ther* 2003;**34**:535–52. https://doi.org/10.1016/S0005-7894(03)80034-5

31. van Lier PA, Muthén BO, van der Sar RM, Crijnen AA. Preventing disruptive behavior in elementary schoolchildren: impact of a universal classroom-based intervention. *J Consult Clin Psychol* 2004;**72**:467–78. https://doi.org/10.1037/0022-006X.72.3.467

32. Witvliet M, van Lier PA, Cuijpers P, Koot HM. Testing links between childhood positive peer relations and externalizing outcomes through a randomized controlled intervention study. *J Consult Clin Psychol* 2009;**77**:905–15. https://doi.org/10.1037/a0014597

33. Hansen WB, Bishop DC, Jackson-Newsom J. Impact of a classroom behavior management intervention on teacher risk ratings for student behavior. *J Drug Educ* 2010;**40**:81–90. https://doi.org/10.2190/DE.40.1.f

34. Leflot G, van Lier PA, Onghena P, Colpin H. The role of teacher behavior management in the development of disruptive behaviors: an intervention study with the good behavior game. *J Abnorm Child Psychol* 2010;**38**:869–82. https://doi.org/10.1007/s10802-010-9411-4

35. Dion E, Roux C, Landry D, Fuchs D, Wehby J, Dupéré V. Improving attention and preventing reading difficulties among low-income first-graders: a randomized study. *Prev Sci* 2011;**12**:70–9. https://doi.org/10.1007/s11121-010-0182-5

36. O'Keeffe J. *A Feasibility Study and a Pilot Cluster Randomised Controlled Trial of the PAX 'Good Behaviour Game' in Disadvantaged Schools.* Belfast: Queen's University Belfast; 2019.

37. Ialongo NS, Domitrovich C, Embry D, Greenberg M, Lawson A, Becker KD, Bradshaw C. A randomized controlled trial of the combination of two school-based universal preventive interventions. *Dev Psychol* 2019;**55**:1313–25. https://doi.org/10.1037/dev0000715

38. Reid JB, Eddy JM, Fetrow RA, Stoolmiller M. Description and immediate impacts of a preventive intervention for conduct problems. *Am J Community Psychol* 1999;**27**:483–517. https://doi.org/10.1023/A:1022181111368

39. Gupta SK. Intention-to-treat concept: a review. *Perspect Clin Res* 2011;**2**:109–12. https://doi.org/10.4103/2229-3485.83221

40. Domitrovich CE, Gest SD, Jones D, Gill S, Sanford Derousie RM. Implementation quality: lessons learned in the context of the Head Start REDI trial. *Early Child Res Q* 2010;**25**:284–98. https://doi.org/10.1016/j.ecresq.2010.04.001

41. Smith S, Barajas K, Ellis B, Moore C, McCauley S, Reichow B. A meta-analytic review of randomized controlled trials of the Good Behavior Game. *Behav Modif* 2021;**45**:641–66. https://doi.org/10.1177/0145445519878670

42. Tanner-Smith EE, Durlak JA, Marx RA. Empirically based mean effect size distributions for universal prevention programs targeting school-aged youth: a review of meta-analyses. *Prev Sci* 2018;**19**:1091–101. https://doi.org/10.1007/s11121-018-0942-1

43. Farrell AD, Henry DB, Bettencourt A. Methodological challenges examining subgroup differences: examples from universal school-based youth violence prevention trials. *Prev Sci* 2013;**14**:121–33. https://doi.org/10.1007/s11121-011-0200-2

44. Gneezy U, Leonard KL, List JA. Gender differences in competition: evidence from a matrilineal and a patriarchal society. *Econometrica* 2009;**77**:1637–64. https://doi.org/10.3982/ECTA6690

45. Durlak JA. Studying program implementation is not easy but it is essential. *Prev Sci* 2015;**16**:1123–7. https://doi.org/10.1007/s11121-015-0606-3

46. Durlak JA. Programme implementation in social and emotional learning: basic issues and research findings. *Cambridge J Educ* 2016;**46**:333–45. https://doi.org/10.1080/0305764X.2016.1142504

47. Bradshaw CP, Shukla KD, Pas ET, Berg JK, Ialongo NS. Using complier average causal effect estimation to examine student outcomes of the PAX Good Behavior Game when integrated with the PATHS curriculum. *Adm Policy Ment Health* 2020;**47**:972–86. https://doi.org/10.1007/s10488-020-01034-1

48. Ashworth E, Panayiotou M, Humphrey N, Hennessey A. Game on – complier average causal effect estimation reveals sleeper effects on academic attainment in a randomized trial of the Good Behavior Game. *Prev Sci* 2020;**21**:222–33. https://doi.org/10.1007/s11121-019-01074-6

49. Kellam SG, Wang W, Mackenzie AC, Brown CH, Ompad DC, Or F, *et al.* The impact of the Good Behavior Game, a universal classroom-based preventive intervention in first and second grades, on high-risk sexual behaviors and drug abuse and dependence disorders into young adulthood. *Prev Sci* 2014;**15**(Suppl. 1):6–18. https://doi.org/10.1007/s11121-012-0296-z

50. Phillips D, Christie F. Behaviour management in a secondary school classroom: playing the game. *Mal Ther Educ* 1986;**4**:47–53.

51. Webster JB. Applying behavior management principles with limited resources: going it alone. *Mal Ther Educ* 1989;**7**:30–8.

52. Chan G, Foxcroft D, Smurthwaite B, Coombes L, Allen D. *Improving Child Behaviour Management: An Evaluation of the Good Behaviour Game in UK Primary Schools.* Oxford: Oxford Brookes University; 2012.

53. Ford C, Keegan N, Poduska J, Kellam S, Littman J. *Good Behaviour Game Implementation Manual.* Washington, DC: American Institutes for Research; 2014.

54. Lendrum A, Humphrey N. The importance of studying the implementation of school-based interventions. *Oxford Rev Educ* 2012;**38**:635–52. https://doi.org/10.1080/03054985.2012.734800

55. Masten AS, Cicchetti D. Developmental cascades. *Dev Psychopathol* 2010;**22**:491–5. https://doi.org/10.1017/S0954579410000222

56. Moilanen KL, Shaw DS, Maxwell KL. Developmental cascades: externalizing, internalizing, and academic competence from middle childhood to early adolescence. *Dev Psychopathol* 2010;**22**:635–53. https://doi.org/10.1017/S0954579410000337

57. Poduska JM, Kellam SG, Wang W, Brown CH, Ialongo NS, Toyinbo P. Impact of the Good Behavior Game, a universal classroom-based behavior intervention, on young adult service use for problems with emotions, behavior, or drugs or alcohol. *Drug Alcohol Depend* 2008;**95**(Suppl. 1):S29–44. https://doi.org/10.1016/j.drugalcdep.2007.10.009

58. Puffer S, Torgerson DJ, Watson J. Cluster randomized controlled trials. *J Eval Clin Pract* 2005;**11**:479–83. https://doi.org/10.1111/j.1365-2753.2005.00568.x

59. Treasure T, MacRae KD. Minimisation: the platinum standard for trials? Randomisation doesn't guarantee similarity of groups; minimisation does. *BMJ* 1998;**317**:362–3. https://doi.org/10.1136/bmj.317.7155.362

60. Goodman R. The Strengths and Difficulties Questionnaire: a research note. *J Child Psychol Psychiatry* 1997;**38**:581–6. https://doi.org/10.1111/j.1469-7610.1997.tb01545.x

61. Department for Education. *Find and Compare Schools in England*. URL: www.compare-school-performance.service.gov.uk/ (accessed July 2015).

62. Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D, *et al.* Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. *BMJ* 2014;**348**:g1687. https://doi.org/10.1136/bmj.g1687

63. Humphrey N, Lendrum A, Ashworth E, Frearson K, Buck R, Kerr K. *Implementation and Process Evaluation (IPE) for Interventions in Educational Settings: an Introductory Handbook*. London: Education Endowment Foundation; 2016.

64. Elswick S, Casey L. The Good Behavior Game is no longer just an effective intervention for students: an examination of the reciprocal effects on teacher behaviors. *Beyond Behav* 2011;**21**:36–46.

65. Durlak JA, DuPre EP. Implementation matters: a review of research on the influence of implementation on program outcomes and the factors affecting implementation. *Am J Community Psychol* 2008;**41**:327–50. https://doi.org/10.1007/s10464-008-9165-0

66. US Department of Health and Human Services. *Finding the Balance: Program Fidelity and Adaptation in Substance Abuse Prevention*. Washington, DC: US Department of Health and Human Services; 2002.

67. Moore JE, Bumbarger BK, Cooper BR. Examining adaptations of evidence-based programs in natural contexts. *J Prim Prev* 2013;**34**:147–61. https://doi.org/10.1007/s10935-013-0303-6

68. Reupert A, Woodcock S. Success and near misses: pre-service teachers' use, confidence and success in various classroom management strategies. *Teach Teach Educ* 2010;**26**:1261–8. https://doi.org/10.1016/j.tate.2010.03.003

69. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, *et al.* Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol* 2007;**60**:34–42. https://doi.org/10.1016/j.jclinepi.2006.03.012

70. The KIDSCREEN Group Europe. *The KIDSCREEN Questionnaires – Quality of life questionnaires for children and adolescents*. Lengerich: Pabst Science Publishers; 2006.

71. Department for Education. *National Pupil Database*. URL: www.gov.uk/government/collections/national-pupil-database (accessed November 2020).

72. Vincent D, Krumpler M. *Hodder Group Reading Test Manual*. London: Hodder; 2007.

73. Koth CW, Bradshaw CP, Leaf PJ. Teacher observation of classroom adaptation–checklist: development and factor structure. *Meas Eval Couns Dev* 2009;**42**:15–30. https://doi.org/10.1177/0748175609333560

74. Warren SF, Fey ME, Yoder PJ. Differential treatment intensity research: a missing link to creating optimally effective communication interventions. *Ment Retard Dev Disabil Res Rev* 2007;**13**:70–7. https://doi.org/10.1002/mrdd.20139

75. Humphrey N, Lendrum A, Bohnke J, Wigelsworth M, Phillips C, Foxcroft D. *Universal School-Based Prevention: Examining the Impact of the Good Behaviour Game on Health-Related Outcomes for Children*. URL: www.fundingawards.nihr.ac.uk/award/14/52/38 (accessed November 2021).

76. Rosseel Y. lavaan: an R package for structural equation modeling. *J Stat Softw* 2012;**48**:1–36. https://doi.org/10.18637/jss.v048.i02

77. Hallquist MN, Wiley JF. MplusAutomation: an R package for facilitating large-scale latent variable analyses in Mplus. *Struct Equ Model* 2018;**25**:621–38. https://doi.org/10.1080/10705511.2017.1402334

78. Campbell MK, Elbourne DR, Altman DG, CONSORT CLUSTER group. [The CONSORT statement for cluster randomised trials.] *Med Clin* 2005;**125**(Suppl. 1):28–31. https://doi.org/10.1016/S0025-7753(05)72206-5

79. Tymms P. Effect Sizes in Multilevel Models. In Schegen I, Elliot K, editors. *But What Does It Mean? The Use of Effect Sizes in Educational Research.* London: National Foundation for Eduacational Research; 2004. pp. 55–65.

80. Fritz CO, Morris PE, Richler JJ. Effect size estimates: current use, calculations, and interpretation. *J Exp Psychol Gen* 2012;**141**:2–18. https://doi.org/10.1037/a0024338

81. Jo B, Asparouhov T, Muthén BO, Ialongo NS, Brown CH. Cluster randomized trials with treatment noncompliance. *Psychol Methods* 2008;**13**:1–18. https://doi.org/10.1037/1082-989X.13.1.1

82. Berg JK, Bradshaw CP, Jo B, Ialongo NS. Using complier average causal effect estimation to determine the impacts of the Good Behavior Game preventive intervention on teacher implementers. *Adm Policy Ment Health* 2017;**44**:558–71. https://doi.org/10.1007/s10488-016-0738-1

83. LeBreton JM, Senter JL. Answers to 20 questions about interrater reliability and interrater agreement. *Organ Res Methods* 2008;**11**:815–52. https://doi.org/10.1177/1094428106296642

84. Muthén L, Muthén BO. *MPlus User's Guide.* Los Angeles, CA: Muthén & Muthén; 2018.

85. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. *J Am Stat Assoc* 1996;**91**:444. https://doi.org/10.1080/01621459.1996.10476902

86. Hamaker EL, Kuiper RM, Grasman RP. A critique of the cross-lagged panel model. *Psychol Methods* 2015;**20**:102–16. https://doi.org/10.1037/a0038889

87. Office for National Statistics. *Disclosure Control: Best Practice for Applying Digital Control to Data.* URL: www.ons.gov.uk/methodology/methodologytopicsandstatisticalconcepts/disclosurecontrol (accessed November 2021).

88. Skrondal A, Rabe-Hesketh S. *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models.* London: Chapman & Hall; 2004. https://doi.org/10.1201/9780203489437

89. Troncoso P, Morales-Gómez A. Estimating the complier average causal effect via a latent class approach using gsem. *Stata J* 2022; in press.

90. Little RJ, Yau LHY. Statistical techniques for analyzing data from prevention trials: treatment of no-shows using Rubin's causal model. *Psychol Methods* 1998;**3**:147–59. https://doi.org/10.1037/1082-989X.3.2.147

91. Muthén BO, Muthén L. *Complier Average Causal Effect (CACE) Estimation in a Randomized Trial.* Los Angeles, CA: Muthén & Muthén; 2010.

92. Vermunt JK, Magidson J. *Technical Guide for Latent GOLD 5.1: Basic, Advanced, and Syntax.* Belmont, MA: Statistical Innovations Inc.; 2016.

93. Bryk A, Raudenbush S. Application of hierarchical linear models to assessing change. *Psychol Bull* 1987;**101**:147–58. https://doi.org/10.1037/0033-2909.101.1.147

94. Goldstein H. *Multilevel Models in Educational and Social Research.* London/New York, NY: Griffin and Oxford Press; 1987.

95. Goldstein H. Models for Multilevel Response Variables with an Application to Growth Curves. In Bock R, editor. *Multilevel Analysis of Educational Data*. San Diego, CA: Academic Press; 1989. pp. 107–25. https://doi.org/10.1016/B978-0-12-108840-8.50011-1

96. Ford T, Hayes R, Byford S, Edwards V, Fletcher M, Logan S, *et al.* The effectiveness and cost-effectiveness of the Incredible Years® Teacher Classroom Management programme in primary school children: results of the STARS cluster randomised controlled trial. *Psychol Med* 2019;**49**:828–42. https://doi.org/10.1017/S0033291718001484

97. Humphrey N. *GBG Trial Interview*. [YouTube]. 17 March 2021. URL: https://youtu.be/vIP3VBs1_R8

98. Department for Education. *Schools, Pupils and Their Characteristics: January 2016*. London: Department for Education; 2016.

99. Department for Education. *Pupil Absence in Schools in England: 2014 to 2015*. London: Department for Education; 2016.

100. Department for Education. *Schools, Pupils and Their Characteristics: January 2015*. London: Department for Education; 2015.

101. Department for Education. *Special Educational Needs in England: January 2015*. London: Department for Education; 2015.

102. Department for Education. *National Curriculum Assessments at Key Stage 2 in England, 2015 (Revised)*. London: Department for Education; 2015.

103. Grimm K, Ram N, Estabrook R. *Growth Modeling: Structural Equation and Multilevel Modeling Approaches*. New York, NY/London: The Guilford Press; 2017.

104. Jung T, Wickrama KAS. An introduction to latent class growth analysis and growth mixture modeling. *Soc Personal Psychol Compass* 2008;**2**:302–17. https://doi.org/10.1111/j.1751-9004.2007.00054.x

105. Romeo R, Knapp M, Scott S. Economic cost of severe antisocial behaviour in children – and who pays it. *Br J Psychiatry* 2006;**188**:547–53. https://doi.org/10.1192/bjp.bp.104.007625

106. National Institute of Health and Care Excellence (NICE). *Methods for the Development of NICE Public Health Guidance (Third Edition)*. London: NICE; 2012.

107. Bell K, Corbacho B, Ronaldson S, Richardson G, Hood K, Sanders J, *et al.* Costs and consequences of the Family Nurse Partnership (FNP) programme in England: evidence from the Building Blocks trial. *F1000Res* 2019;**8**:1640. https://doi.org/10.12688/f1000research.20149.1

108. Office for National Statistics (ONS). *Employee Earnings in the UK Statistical Bulletins*. London: ONS; 2020.

109. Office for National Statistics (ONS). *Annual survey of hours and earnings: 2015 provisional results*. Newport: ONS; 2015.

110. Office for National Statistics (ONS). *Annual survey of hours and earnings: 2016 provisional results*. Newport: ONS; 2016.

111. Office for National Statistics (ONS). *Annual survey of hours and earnings: 2017 provisional and 2016 revised results*. Newport: ONS; 2017.

112. Groves EA, Austin JL. Does the Good Behavior Game evoke negative peer pressure? Analyses in primary and secondary classrooms. *J Appl Behav Anal* 2019;**52**:3–16. https://doi.org/10.1002/jaba.513

113. Greenberg MT. School-based prevention: current status and future challenges. *Eff Educ* 2010;**2**:27–52. https://doi.org/10.1080/19415531003616862

114. Elswick S, Casey LB, Zanskas S, Black T, Schnell R. Effective data collection modalities utilized in monitoring the good behavior game: technology-based data collection versus hand collected data. *Comput Human Behav* 2016;**54**:158–69. https://doi.org/10.1016/j.chb.2015.07.059

115. Liu J. Minimum Effective Dose. In Chow S-C, editor. *Encyclopedia of Biopharmaceutical Statistics*. London: Informa PLC; 2010. pp. 799–800. https://doi.org/10.3109/9781439822463.128

116. Pennington B, McComas JJ. Effects of the Good Behavior Game across classroom contexts. *J Appl Behav Anal* 2017;**50**:176–80. https://doi.org/10.1002/jaba.357

117. Domitrovich CE, Bradshaw CP, Greenberg MT, Embry D, Poduska JM, Ialongo NS. Intergrated models of school-based prevention: logic and theory. *J Adolesc* 2010;**47**:71–88. https://doi.org/10.1002/pits.20452

118. Hill KG, Woodward D, Woelfel T, Hawkins JD, Green S. Planning for long-term follow-up: strategies learned from longitudinal studies. *Prev Sci* 2016;**17**:806–18. https://doi.org/10.1007/s11121-015-0610-7

119. Greenberg MT, Abenavoli R. Universal interventions: fully exploring their impacts and potential to produce population-level impacts. *J Res Educ Eff* 2017;**10**:40–67. https://doi.org/10.1080/19345747.2016.1246632

120. Washington State Institute for Public Policy. *Good Behavior Game Benefit-Cost Estimates*. Olympia, WA: Washington State Institute for Public Policy; 2019.

121. Bonin EM, Stevens M, Beecham J, Byford S, Parsonage M. Costs and longer-term savings of parenting programmes for the prevention of persistent conduct disorder: a modelling study. *BMC Public Health* 2011;**11**:803. https://doi.org/10.1186/1471-2458-11-803

122. Agus A, McKay M, Cole J, Doherty P, Foxcroft D, Harvey S, *et al.* Cost-effectiveness of a combined classroom curriculum and parental intervention: economic evaluation of data from the Steps Towards Alcohol Misuse Prevention Programme cluster randomised controlled trial. *BMJ Open* 2019;**9**:e027951. https://doi.org/10.1136/bmjopen-2018-027951

123. Canaway A, Frew E, Lancashire E, Pallan M, Hemming K, Adab P, WAVES trial investigators. Economic evaluation of a childhood obesity prevention programme for children: results from the WAVES cluster randomised controlled trial conducted in schools. *PLOS ONE* 2019;**14**:e0219500. https://doi.org/10.1371/journal.pone.0219500

124. Clemes SA, Bingham DD, Pearson N, Chen Y-L, Edwardson C, McEachan R, *et al.* Sit–stand desks to reduce sedentary behaviour in 9- to 10-year-olds: the Stand Out in Class pilot cluster RCT. *Public Heal Res* 2020;**8**:1–126. https://doi.org/10.3310/phr08080

125. Masten AS, Roisman GI, Long JD, Burt KB, Obradović J, Riley JR, *et al.* Developmental cascades: linking academic achievement and externalizing and internalizing symptoms over 20 years. *Dev Psychol* 2005;**41**:733–46. https://doi.org/10.1037/0012-1649.41.5.733

126. Gorard S. The propagation of errors in experimental data analysis: a comparison of pre- and post-test designs. *Int J Res Method Educ* 2013;**36**:372–85. https://doi.org/10.1080/1743727X.2012.741117

127. Education Endowment Foundation. *Improving Behaviour in Schools: Guidance Report*. London: Education Endowment Foundation; 2019.

128. Moore D, Benham-Clarke S, Kenchington R, Boyle C, Ford T, Hayes R, *et al. Improving Behaviour in Schools: Evidence Review*. London: Education Endowment Foundation; 2019.

129. Humphrey N, Panayiotou M, Hennessey A, Ashworth E. Treatment effect modifiers in a randomized trial of the good behavior game during middle childhood. *J Consult Clin Psychol* 2021;**89**:668–81. https://doi.org/10.1037/ccp0000673

130. Troncoso P, Humphrey N. Playing the long game: a multivariate multilevel non-linear growth curve model of long-term effects in a randomized trial of the Good Behaviour Game. *Journ of School Psy* 2021;**88**:66–84. https://doi.org/10.1016/j.jsp.2021.08.002

131. Burns RA, Crisp DA, Burns RB. Re-examining the reciprocal effects model of self-concept, self-efficacy, and academic achievement in a comparison of the Cross-Lagged Panel and Random-Intercept Cross-Lagged Panel frameworks. *Br J Educ Psychol* 2020;**90**:77–91. https://doi.org/10.1111/bjep.12265

Appendix 1 Additional full model information for hypothesis 3

This section describes the full CACE models for hypothesis 3. Specifically, each of the following seven tables describe the moderate and high CACE compliance findings for each of the seven outcome variables. All tables summarise the sample size within each class, the effect of key student-level and school-level covariates for each outcome beyond the intervention effect, and the total variance predicted by these (R^2). These coefficients are available for each of the four classes of individuals: moderate compliers versus non-compliers, and high compliers versus non-compliers.

The models of *Tables 28–33* were analysed in M*plus* 8.4 and results are presented in standardised form. This allows for a direct comparison of effects between models. The entropy of these models is presented at the bottom of each table and indicates the delineation of classes, with higher values indicating better classification of individuals in the two classes (e.g. moderate compliers vs. non-compliers).

TABLE 28 Full CACE model for conduct problems (moderate and high compliance)

Variable	CACE standardised coefficient (SE)			
	Moderate compliance		**High compliance**	
	Compliers	**Non-compliers**	**Compliers**	**Non-compliers**
Number	1378	1276	701	1954
Child level (R^2)	0.51***	0.36***	0.62***	0.37***
Baseline (T1)	0.448 (0.049)***	0.3 (0.05)***	0.473 (0.078)***	0.34 (0.042)***
KS1 attainment	–0.077 (0.051)	0.098 (0.053)	–0.011 (0.082)	–0.002 (0.048)
Concentration problems	0.25 (0.067)***	0.119 (0.076)	0.324 (0.091)***	0.163 (0.056)**
Prosocial behaviour	–0.059 (0.054)	–0.081 (0.063)	–0.084 (0.074)	–0.067 (0.054)
FSMs (0 = no; 1 = yes)	0.220 (0.095)*	–0.002 (0.138)	0.244 (0.153)	0.074 (0.102)
Special educational needs and disabilities	–0.064 (0.116)	0.427 (0.158)**	0.098 (0.154)	0.163 (0.117)
Sex (1 = male; 2 = female)	–0.024 (0.119)	0.541 (0.137)***	–0.214 (0.164)	0.384 (0.114)**
School level (R^2)	0.36**	0.27*	0.44	0.28
Trial (1 = usual practice; 2 = GBG)	0.006 (0.248)	–	0.258 (0.539)	–
School size	–0.063 (0.085)	0.057 (0.265)	–0.481 (0.623)	0.003 (0.074)
FSMs (%)	0.047 (0.144)	0.322 (0.213)	–0.264 (0.186)	0.224 (0.132)
School conduct problems	–0.455 (0.093)***	0.242 (0.19)	–0.427 (0.283)	–0.295 (0.165)
English as additional language (%)	–0.368 (0.122)**	0.138 (0.255)	0.348 (0.558)	–0.109 (0.135)
KS1 school attainment	0.076 (0.178)	0.396 (0.214)	–0.417 (0.192)*	0.447 (0.126)

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

Notes
Entropy moderate = 0.79; entropy high = 0.76. M*plus* 8.4 was used for these models.

TABLE 29 Full CACE model for psychological well-being (moderate and high compliance)

	CACE standardised coefficient (SE)			
	Moderate compliance		High compliance	
Variable	Compliers	Non-compliers	Compliers	Non-compliers
Number	1057	1566	616	2007
Child level (R^2)	0.02	0.05**	0.02	0.05**
Conduct problems T1	0.016 (0.056)	−0.143 (0.04)***	0.019 (0.078)	−0.103 (0.035)**
KS1 attainment	−0.042 (0.069)	0.111 (0.05)*	−0.16 (0.081)	0.109 (0.045)*
Concentration problems	−0.088 (0.073)	0.002 (0.05)	−0.16 (0.108)	−0.011 (0.047)
Prosocial behaviour	0.048 (0.08)	0.056 (0.044)	−0.042 (0.098)	0.083 (0.036)*
FSMs (0 = no; 1 = yes)	−0.077 (0.089)	−0.053 (0.055)	−0.12 (0.144)	−0.045 (0.058)
Special educational needs and disabilities	−0.105 (0.134)	0.171 (0.081)*	−0.072 (0.139)	0.107 (0.068)
Sex (1 = male; 2 = female)	−0.056 (0.07)	0.046 (0.058)	−0.088 (0.109)	0.039 (0.049)
School level (R^2)	0.84***	0.54*	0.76***	0.18
Trial (1 = usual practice; 2 = GBG)	1.239 (0.377)**	–	−0.959 (0.38)*	–
School size	0.06 (0.414)	−0.251 (0.169)	−0.033 (0.416)	−0.182 (0.13)
FSMs (%)	0.385 (0.227)	−0.647 (0.265)*	0.795 (0.217)***	−0.401 (0.27)
School conduct problems	−0.42 (0.152)**	0.315 (0.369)	−0.139 (0.24)	−0.023 (0.183)
English as additional language (%)	−0.266 (0.307)	0.303 (0.27)	−0.144 (0.258)	0.106 (0.164)
KS1 school attainment	−0.145 (0.191)	−0.196 (0.275)	0.385 (0.175)*	−0.277 (0.195)

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

Notes
Entropy moderate = 0.78; entropy high = 0.84. M*plus* 8.4 was used for these models.

TABLE 30 Full CACE model for emotional symptoms (moderate and high compliance)

	CACE standardised coefficient (SE)			
	Moderate compliance		High compliance	
Variable	Compliers	Non-compliers	Compliers	Non-compliers
Number	1389	1266	775	1880
Child level (R^2)	0.10**	0.08*	0.13***	0.08**
Conduct problems T1	0.067 (0.18)	0.085 (0.252)	0.165 (0.09)	0.012 (0.075)
KS1 attainment	−0.183 (0.048)***	0.025 (0.093)	−0.213 (0.046)***	−0.063 (0.061)
Concentration problems	−0.182 (0.151)	0.099 (0.206)	−0.305 (0.099)**	0.056 (0.094)
Prosocial behaviour	−0.248 (0.056)***	0.005 (0.103)	−0.287 (0.077)***	−0.037 (0.072)
FSMs (0 = no; 1 = yes)	0.186 (0.148)	0.016 (0.225)	0.17 (0.108)	0.071 (0.109)
Special educational needs and disabilities	−0.006 (0.173)	0.484 (0.287)	−0.061 (0.189)	0.405 (0.199)*
Sex (1 = male; 2 = female)	−0.278 (0.105)**	−0.177 (0.172)	−0.288 (0.096)**	−0.201 (0.08)*
School level (R^2)	0.997***	0.995***	0.62***	0.62***

TABLE 30 Full CACE model for emotional symptoms (moderate and high compliance) (*continued*)

Variable	CACE standardised coefficient (SE)			
	Moderate compliance		High compliance	
	Compliers	Non-compliers	Compliers	Non-compliers
Trial (1 = usual practice; 2 = GBG)	–0.247 (0.38)	–	–0.341 (0.428)	–
School size	–0.377 (0.319)	0.429 (0.175)*	–0.509 (0.312)	0.267 (0.103)*
FSMs (%)	–0.13 (0.238)	–0.089 (0.377)	–0.021 (0.178)	–0.457 (0.238)
School conduct problems	0.182 (0.178)	0.664 (0.351)	–0.098 (0.173)	0.517 (0.253)*
English as additional language (%)	–0.679 (0.271)*	0.197 (0.218)	–0.306 (0.404)	–0.162 (0.177)
KS1 school attainment	–0.262 (0.192)	0.593 (0.451)	–0.345 (0.137)*	0.248 (0.251)

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

Notes
Entropy moderate = 0.73; entropy high = 0.86. M*plus* 8.4 was used for these models.

TABLE 31 Full CACE model for peer and social support (moderate and high compliance)

Variable	CACE standardised coefficient (SE)			
	Moderate compliance		High compliance	
	Compliers	Non-compliers	Compliers	Non-compliers
Number	1217	1425	629	2013
Child level (R^2)	0.05*	0.03	0.05	0.03
Conduct problems T1	0.164 (0.058)**	–0.104 (0.042)*	0.084 (0.639)	–0.004 (0.166)
KS1 attainment	–0.143 (0.055)**	–0.003 (0.045)	–0.183 (0.487)	–0.04 (0.202)
Concentration problems	–0.088 (0.058)	0.067 (0.054)	–0.145 (0.325)	0.021 (0.062)
Prosocial behaviour	0.161 (0.069)*	0.111 (0.069)	–0.013 (0.295)	0.155 (0.091)
FSMs (0 = no; 1 = yes)	–0.135 (0.097)	0.168 (0.096)	–0.345 (0.14)*	0.104 (0.092)
Special educational needs and disabilities	–0.365 (0.134)**	–0.019 (0.098)	–0.165 (2)	–0.164 (0.659)
Sex (1 = male; 2 = female)	–0.064 (0.069)	–0.039 (0.076)	–0.219 (0.383)	–0.005 (0.19)
School level (R^2)	0.75***	0.29	0.80	0.08
Trial (1 = usual practice; 2 = GBG)	0.491 (0.341)	–	–0.246 (2.341)	–
School size	0.937 (0.151)***	–0.402 (0.151)**	0.783 (1.204)	–0.124 (0.221)
FSMs (%)	0.185 (0.205)	0.346 (0.393)	0.629 (0.232)**	0.037 (0.476)
School conduct problems	0.352 (0.201)	0.097 (0.549)	0.307 (0.999)	0.194 (0.333)
English as additional language (%)	–0.678 (0.152)***	0.196 (0.355)	–0.563 (1.471)	0.182 (1.182)
KS1 school attainment	0.184 (0.174)	–0.098 (0.328)	0.301 (0.789)	–0.094 (0.39)

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

Notes
Entropy moderate = 0.73; entropy high = 0.85. M*plus* 8.4 was used for these models.

TABLE 32 Full CACE model for school environment (moderate and high compliance)

	CACE standardised coefficient (SE)			
	Moderate compliance		High compliance	
Variable	Compliers	Non-compliers	Compliers	Non-compliers
Number	1271	1365	601	2030
Student-level (R^2)	0.09***	0.13**	0.13***	0.10***
Conduct problems T1	−0.047 (0.074)	−0.202 (0.062)**	0.074 (0.065)	−0.184 (0.047)***
KS1 attainment	0 (0.056)	−0.003 (0.042)	−0.123 (0.067)	0.023 (0.042)
Concentration problems	−0.232 (0.064)***	−0.115 (0.051)*	−0.353 (0.055)***	−0.131 (0.051)*
Prosocial behaviour	−0.038 (0.086)	0.096 (0.057)	0.129 (0.085)	−0.012 (0.047)
FSMs (0 = no; 1 = yes)	0.046 (0.096)	0.134 (0.077)	0.127 (0.128)	0.073 (0.061)
Special educational needs and disabilities	−0.048 (0.1)	0.138 (0.084)	0.114 (0.201)	0.062 (0.073)
Sex (1 = male; 2 = female)	−0.231 (0.074)**	−0.169 (0.061)**	−0.062 (0.151)	−0.239 (0.045)***
School level (R^2)	0.35	0.49*	0.65**	0.16
Trial (1 = usual practice; 2 = GBG)	−0.121 (0.671)	–	0.044 (1.224)	–
School size	0.543 (0.329)	−0.193 (0.132)	0.545 (0.21)**	−0.184 (0.162)
FSMs (%)	0.142 (0.238)	0.175 (0.242)	0.312 (0.274)	0.003 (0.292)
School conduct problems	−0.117 (0.183)	0.594 (0.231)*	−0.112 (0.379)	0.232 (0.239)
English as additional language (%)	−0.559 (0.215)**	0.208 (0.133)	−0.841 (0.213)***	0.107 (0.157)
KS1 school attainment	−0.033 (0.256)	−0.112 (0.162)	0.323 (0.158)*	−0.266 (0.178)

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

Notes
Entropy moderate = .68; High = .80. M*plus* 8.4 was used for these models.

TABLE 33 Full CACE model for bullying (i.e. social acceptance) (moderate and high compliance)

	CACE standardised coefficient (SE)			
	Moderate compliance		High compliance	
Variable	Compliers	Non-compliers	Compliers	Non-compliers
Number	1395	1244	744	1895
Child level (R^2)	0.05	0.11	0.11***	0.06***
Conduct problems T1	0.043 (0.062)	−0.092 (0.278)	0.113 (0.097)	−0.087 (0.051)
KS1 attainment	0.107 (0.446)	0.254 (0.297)	0.128 (0.086)	0.189 (0.042)***
Concentration problems	−0.059 (0.062)	−0.025 (0.115)	−0.023 (0.086)	−0.041 (0.046)
Prosocial behaviour	0.056 (0.097)	0.036 (0.164)	0.124 (0.116)	0.018 (0.05)
FSMs (0 = no; 1 = yes)	0.051 (0.231)	−0.243 (0.305)	−0.223 (0.15)	−0.025 (0.089)
Special educational needs and disabilities	−0.122 (0.238)	0.216 (0.307)	−0.277 (0.209)	0.16 (0.095)
Sex (1 = male; 2 = female)	0.283 (0.578)	0.139 (0.99)	0.399 (0.101)***	0.128 (0.067)
School level (R^2)	0.67**	0.84	0.23	0.35
Trial (1 = usual practice; 2 = GBG)	0.37 (2.989)	–	0.235 (0.592)	–
School size	−0.65 (1.602)	0.535 (3.25)	−0.233 (0.4)	−0.42 (0.225)

TABLE 33 Full CACE model for bullying (i.e. social acceptance) (moderate and high compliance) (*continued*)

| | CACE standardised coefficient (SE) | | | |
| | Moderate compliance | | High compliance | |
Variable	Compliers	Non-compliers	Compliers	Non-compliers
FSMs (%)	0.425 (0.98)	−0.491 (3.284)	0.281 (0.397)	−0.407 (0.385)
School conduct problems	−0.487 (0.813)	0.612 (1.219)	−0.309 (0.277)	0.19 (0.324)
English as additional language (%)	0.138 (2.029)	−0.098 (4.156)	−0.032 (0.405)	0.386 (0.288)
KS1 school attainment	0.125 (1.187)	0.23 (3.66)	0.23 (0.167)	−0.04 (0.327)

$*p < 0.05$, $**p < 0.01$, $***p < 0.001$.

Notes
Entropy moderate $= 0.73$; entropy high $= 0.77$. M*plus* 8.4 was used for these models.

The model of *Table 34*, which represents a negative binomial CACE, was analysed in Stata, version 16.1, and results are presented in unstandardised form. The model fit coefficients Bayesian information criterion and Akaike information criterion are presented at the bottom of the table, for which models with lower values indicate a better fit.

TABLE 34 Full CACE model for school absence (moderate and high compliance)

| | CACE coefficient (SE[a]) | | | |
| | Moderate compliance | | High compliance | |
Variable	Compliers	Non-compliers	Compliers	Non-compliers
Child level				
Conduct problems T1	0.034 (0.019)	−0.014 (0.016)	0.038 (0.028)	0.0002 (0.022)
KS1 attainment	−0.024 (0.010)*	−0.032 (0.016)*	−0.043 (0.016)**	−0.024 (0.011)*
Concentration problems	−0.048 (0.032)	−0.024 (0.036)	−0.093 (0.048)	−0.026 (0.027)
Prosocial behaviour	−0.051 (0.042)	0.032 (0.047)	−0.020 (0.064)	0.010 (0.056)
FSMs (0 = no; 1 = yes)	0.041 (0.072)	0.251 (0.102)*	0.063 (0.112)	0.166 (0.076)*
Special educational needs and disabilities	−0.133 (0.089)	0.173 (0.112)	−0.107 (0.107)	0.137 (0.087)
Sex (1 = male; 2 = female)	−0.009 (0.051)	−0.015 (0.064)	0.051 (0.069)	−0.010 (0.063)
Baseline school absence	3.560 (0.542)***	13.298 (1.316)***	3.040 (0.601)***	11.92 (0.946)***
School level				
Trial (1 = usual practice; 2 = GBG)	−0.656 (0.072)***	–	−0.674 (0.162)***	–
School size	−0.0001 (0.0002)	−0.00003 (0.0003)	−0.0003 (0.0004)	−0.0002 (0.0003)
FSMs (%)	0.009 (0.003)*	−0.002 (0.005)	0.007 (0.005)	0.0002 (0.004)
School conduct problems	−0.054 (0.067)	0.105 (0.103)	−0.191 (0.186)	0.101 (0.089)
English as additional language (%)	−0.001 (0.002)	0.003 (0.003)	−0.001 (0.002)	0.003 (0.002)
KS1 school attainment	0.002 (0.004)	0.005 (0.005)	0.002 (0.005)	0.003 (0.004)

$*p < 0.05$, $**p < 0.01$, $***p < 0.001$.
a   Robust SEs clustered by school.

Note
$N = 2888$. Akaike information criterion moderate $= 35977.64$, Bayesian information criterion moderate $= 36234.28$, Akaike information criterion high $= 35836.43$, Bayesian information criterion high $= 36093.07$. Stata, version 16.1, was used for these models.

Appendix 2 Additional full model information for hypothesis 5

TABLE 35 Baseline model (hypothesis 0): measurement part 1

Variable	Coefficient	SE	*p*-value	Standardised coefficient
Conduct problems T3 measured by				
sdq5_3	1.258	0.068	0.000	0.872
sdq7_3	0.687	0.029	0.000	0.697
sdq12_3	1.358	0.075	0.000	0.887
sdq18_3	1.008	0.045	0.000	0.819
sdq22_3	0.713	0.076	0.000	0.710
Conduct problems T4 measured by				
sdq5_4	1.258	0.068	0.000	0.887
sdq7_4	0.687	0.029	0.000	0.723
sdq12_4	1.358	0.075	0.000	0.901
sdq18_4	1.008	0.045	0.000	0.838
sdq22_4	0.713	0.076	0.000	0.736
Conduct problems T5 measured by				
sdq5_5	1.258	0.068	0.000	0.891
sdq7_5	0.687	0.029	0.000	0.732
sdq12_5	1.358	0.075	0.000	0.905
sdq18_5	1.008	0.045	0.000	0.845
sdq22_5	0.713	0.076	0.000	0.745
Emotional symptoms T3 measured by				
sdq3_3	0.589	0.028	0.000	0.640
sdq8_3	1.551	0.052	0.000	0.910
sdq13_3	1.130	0.048	0.000	0.848
sdq16_3	1.070	0.035	0.000	0.834
sdq24_3	1.740	0.077	0.000	0.926
Emotional symptoms T4 measured by				
sdq3_4	0.589	0.028	0.000	0.643
sdq8_4	1.551	0.052	0.000	0.911
sdq13_4	1.130	0.048	0.000	0.850
sdq16_4	1.070	0.035	0.000	0.836
sdq24_4	1.740	0.077	0.000	0.927
Emotional symptoms T5 measured by				
sdq3_5	0.589	0.028	0.000	0.640
sdq8_5	1.551	0.052	0.000	0.910
sdq13_5	1.130	0.048	0.000	0.848
sdq16_5	1.070	0.035	0.000	0.835
sdq24_5	1.740	0.077	0.000	0.927

TABLE 35 Baseline model (hypothesis 0): measurement part 1 (*continued*)

Variable	Coefficient	SE	*p*-value	Standardised coefficient
Thresholds (constrained to equality over time)				
sdq5 threshold 1	3.184	0.183	0.000	1.560
sdq5 threshold 2	4.847	0.244	0.000	2.375
sdq7 threshold 1	1.507	0.071	0.000	1.081
sdq7 threshold 2	3.000	0.094	0.000	2.152
sdq12 threshold 1	3.300	0.190	0.000	1.524
sdq12 threshold 2	5.403	0.273	0.000	2.496
sdq18 threshold 1	2.428	0.124	0.000	1.394
sdq18 threshold 2	4.071	0.169	0.000	2.337
sdq22 threshold 1	3.383	0.243	0.000	2.382
sdq22 threshold 2	4.393	0.270	0.000	3.093
sdq3 threshold 1	1.272	0.062	0.000	0.977
sdq3 threshold 2	2.413	0.076	0.000	1.854
sdq8 threshold 1	1.725	0.114	0.000	0.716
sdq8 threshold 2	4.416	0.167	0.000	1.832
sdq13 threshold 1	2.005	0.114	0.000	1.064
sdq13 threshold 2	4.032	0.162	0.000	2.139
sdq16 threshold 1	1.560	0.091	0.000	0.860
sdq16 threshold 2	3.566	0.124	0.000	1.966
sdq24 threshold 1	2.817	0.174	0.000	1.060
sdq24 threshold 2	5.325	0.254	0.000	2.004
Latent means (intercepts)				
Conduct problems T4	−0.191	0.101	0.058	−0.125
Conduct problems T5	−0.178	0.119	0.135	−0.114
Emotional symptoms T4	−0.031	0.085	0.719	−0.021
Emotional symptoms T5	0.079	0.087	0.362	0.056

TABLE 36 Baseline model (hypothesis 0): measurement part 2

Covariance between observed SDQ items over time	Standardised coefficient	SE	*p*-value
sdq5_3/sdq5_4	0.584	0.104	0.000
sdq5_3/sdq5_5	0.788	0.119	0.000
sdq5_4/sdq5_5	0.460	0.121	0.000
sdq7_3/sdq7_4	0.189	0.048	0.000
sdq7_3/sdq7_5	0.256	0.055	0.000
sdq7_4/sdq7_5	0.159	0.051	0.002
sdq12_3/sdq12_4	0.005	0.120	0.968
sdq12_3/sdq12_5	0.366	0.140	0.009
sdq12_4/sdq12_5	0.338	0.135	0.012
sdq18_3/sdq18_4	0.076	0.098	0.439

TABLE 36 Baseline model (hypothesis 0): measurement part 2 (*continued*)

Covariance between observed SDQ items over time	Standardised coefficient	SE	*p*-value
sdq18_3/sdq18_5	0.245	0.106	0.021
sdq18_4/sdq18_5	0.110	0.101	0.275
sdq22_3/sdq22_4	0.188	0.213	0.377
sdq22_3/sdq22_5	0.371	0.228	0.103
sdq22_4/sdq22_5	0.480	0.169	0.005
sdq3_3/sdq3_4	0.310	0.060	0.000
sdq3_3/sdq3_5	0.252	0.066	0.000
sdq3_4/sdq3_5	0.333	0.061	0.000
sdq8_3/sdq8_4	0.280	0.130	0.032
sdq8_3/sdq8_5	−0.324	0.172	0.060
sdq8_4/sdq8_5	0.087	0.128	0.496
sdq13_3/sdq13_4	0.399	0.104	0.000
sdq13_3/sdq13_5	0.217	0.125	0.081
sdq13_4/sdq13_5	0.176	0.107	0.100
sdq16_3/sdq16_4	0.182	0.096	0.059
sdq16_3/sdq16_5	0.052	0.113	0.646
sdq16_4/sdq16_5	0.419	0.085	0.000
sdq24_3/sdq24_4	0.060	0.199	0.761
sdq24_3/sdq24_5	−0.020	0.262	0.940
sdq24_4/sdq24_5	0.032	0.202	0.876
Scaling factors	**Standardised coefficient**		
sdq5_3	0.490		
sdq7_3	0.717		
sdq12_3	0.462		
sdq18_3	0.574		
sdq22_3	0.704		
sdq5_4	0.462		
sdq7_4	0.691		
sdq12_4	0.435		
sdq18_4	0.545		
sdq22_4	0.677		
sdq5_5	0.453		
sdq7_5	0.681		
sdq12_5	0.426		
sdq18_5	0.535		
sdq22_5	0.668		
sdq3_3	0.769		
sdq8_3	0.415		
sdq13_3	0.531		
sdq16_3	0.551		

TABLE 36 Baseline model (hypothesis 0): measurement part 2 (*continued*)

Scaling factors	Standardised coefficient
sdq24_3	0.376
sdq3_4	0.766
sdq8_4	0.412
sdq13_4	0.528
sdq16_4	0.548
sdq24_4	0.374
sdq3_5	0.768
sdq8_5	0.414
sdq13_5	0.530
sdq16_5	0.551
sdq24_5	0.376

TABLE 37 Baseline model (hypothesis 0): structural part

Cross-lagged effects of within-person factors	Standardised coefficient	SE	*p*-value
Conduct problems T4 regressed on			
Conduct problems T3	0.566	0.081	0.000
Emotional symptoms T3	–0.202	0.082	0.013
Reading attainment T3	0.063	0.042	0.139
Emotional symptoms T4 regressed on			
Conduct problems T3	–0.164	0.138	0.235
Emotional symptoms T3	–0.047	0.066	0.473
Reading attainment T3	0.001	0.029	0.971
Reading attainment T4 regressed on			
Conduct problems T3	0.005	0.053	0.923
Emotional symptoms T3	0.010	0.026	0.695
Reading attainment T3	–0.113	0.011	0.000
Conduct problems T5 regressed on			
Conduct problems T4	0.610	0.077	0.000
Emotional symptoms T4	–0.294	0.071	0.000
Reading attainment T4	0.028	0.025	0.272
Emotional symptoms T5 regressed on			
Conduct problems T4	–0.077	0.119	0.517
Emotional symptoms T4	0.044	0.085	0.609
Reading attainment T4	–0.005	0.021	0.822
Reading attainment T5 regressed on			
Conduct problems T4	–0.014	0.067	0.838
Emotional symptoms T4	–0.005	0.052	0.916
Reading attainment T4	–0.151	0.011	0.000
Covariance within wave of within-person factors			
Conduct problems T3/emotional symptoms T3	0.177	0.126	0.162

TABLE 37 Baseline model (hypothesis 0): structural part (*continued*)

Cross-lagged effects of within-person factors	Standardised coefficient	SE	*p*-value
Conduct problems T3/reading attainment T3	0.028	0.047	0.552
Emotional symptoms T3/reading attainment T3	0.017	0.022	0.454
Conduct problems T4/emotional symptoms T4	0.469	0.078	0.000
Conduct problems T4/reading attainment T4	0.061	0.043	0.154
Emotional symptoms T4/reading attainment T4	0.008	0.031	0.801
Conduct problems T5/emotional symptoms T5	0.422	0.061	0.000
Conduct problems T5/reading attainment T5	0.040	0.038	0.296
Emotional symptoms T5/reading attainment T5	–0.070	0.029	0.017
Covariance between latent random intercepts (between person)			
Conduct problems/emotional symptoms	0.423	0.126	0.001
Conduct problems/reading attainment	–0.224	0.052	0.000
Emotional symptoms/reading attainment	–0.118	0.023	0.000

TABLE 38 Full model (hypothesis 1): measurement part 1

Variable	Coefficient	SE	*p*-value	Standardised coefficient
Conduct problems T3 measured by				
sdq5_3	1.281	0.070	0.000	0.892
sdq7_3	0.702	0.029	0.000	0.734
sdq12_3	1.342	0.071	0.000	0.900
sdq18_3	0.991	0.043	0.000	0.837
sdq22_3	0.725	0.077	0.000	0.745
Conduct problems T4 measured by				
sdq5_4	1.281	0.070	0.000	0.903
sdq7_4	0.702	0.029	0.000	0.755
sdq12_4	1.342	0.071	0.000	0.910
sdq18_4	0.991	0.043	0.000	0.852
sdq22_4	0.725	0.077	0.000	0.765
Conduct problems T5 measured by				
sdq5_5	1.281	0.070	0.000	0.906
sdq7_5	0.702	0.029	0.000	0.762
sdq12_5	1.342	0.071	0.000	0.914
sdq18_5	0.991	0.043	0.000	0.857
sdq22_5	0.725	0.077	0.000	0.772
Emotional symptoms T3 measured by				
sdq3_3	0.592	0.028	0.000	0.653
sdq8_3	1.540	0.051	0.000	0.913
sdq13_3	1.129	0.047	0.000	0.854
sdq16_3	1.074	0.036	0.000	0.842
sdq24_3	1.744	0.077	0.000	0.930

TABLE 38 Full model (hypothesis 1): measurement part 1 (*continued*)

Variable	Coefficient	SE	*p*-value	Standardised coefficient
Emotional symptoms T4 measured by				
sdq3_4	0.592	0.028	0.000	0.655
sdq8_4	1.540	0.051	0.000	0.914
sdq13_4	1.129	0.047	0.000	0.856
sdq16_4	1.074	0.036	0.000	0.844
sdq24_4	1.744	0.077	0.000	0.931
Emotional symptoms T5 measured by				
sdq3_5	0.592	0.028	0.000	0.653
sdq8_5	1.540	0.051	0.000	0.913
sdq13_5	1.129	0.047	0.000	0.855
sdq16_5	1.074	0.036	0.000	0.843
sdq24_5	1.744	0.077	0.000	0.931
Thresholds (constrained to equality over time)				
sdq5 threshold 1	3.229	0.190	0.000	1.459
sdq5 threshold 2	4.913	0.255	0.000	2.220
sdq7 threshold 1	1.525	0.072	0.000	1.035
sdq7 threshold 2	3.033	0.096	0.000	2.059
sdq12 threshold 1	3.270	0.185	0.000	1.423
sdq12 threshold 2	5.350	0.263	0.000	2.328
sdq18 threshold 1	2.401	0.121	0.000	1.315
sdq18 threshold 2	4.023	0.164	0.000	2.203
sdq22 threshold 1	3.413	0.251	0.000	2.275
sdq22 threshold 2	4.431	0.281	0.000	2.954
sdq3 threshold 1	1.275	0.062	0.000	0.966
sdq3 threshold 2	2.419	0.077	0.000	1.832
sdq8 threshold 1	1.716	0.113	0.000	0.699
sdq8 threshold 2	4.390	0.164	0.000	1.789
sdq13 threshold 1	2.005	0.113	0.000	1.042
sdq13 threshold 2	4.030	0.161	0.000	2.095
sdq16 threshold 1	1.564	0.091	0.000	0.843
sdq16 threshold 2	3.576	0.124	0.000	1.927
sdq24 threshold 1	2.821	0.174	0.000	1.035
sdq24 threshold 2	5.332	0.254	0.000	1.955
Latent means (intercepts)				
Conduct problems T4	−0.183	0.098	0.061	−0.112
Conduct problems T5	−0.170	0.115	0.139	−0.102
Emotional symptoms T4	−0.029	0.085	0.730	−0.020
Emotional symptoms T5	0.080	0.087	0.359	0.055

TABLE 39 Full model (hypothesis 1): measurement part 2

Covariance between observed SDQ items over time	Standardised coefficient	SE	p-value
sdq5_3/sdq5_4	0.590	0.107	0.000
sdq5_3/sdq5_5	0.807	0.122	0.000
sdq5_4/sdq5_5	0.460	0.124	0.000
sdq7_3/sdq7_4	0.181	0.049	0.000
sdq7_3/sdq7_5	0.252	0.057	0.000
sdq7_4/sdq7_5	0.149	0.052	0.004
sdq12_3/sdq12_4	0.029	0.117	0.801
sdq12_3/sdq12_5	0.387	0.138	0.005
sdq12_4/sdq12_5	0.354	0.131	0.007
sdq18_3/sdq18_4	0.097	0.095	0.305
sdq18_3/sdq18_5	0.263	0.104	0.011
sdq18_4/sdq18_5	0.131	0.098	0.180
sdq22_3/sdq22_4	0.182	0.217	0.400
sdq22_3/sdq22_5	0.371	0.232	0.110
sdq22_4/sdq22_5	0.479	0.172	0.005
sdq3_3/sdq3_4	0.309	0.060	0.000
sdq3_3/sdq3_5	0.251	0.067	0.000
sdq3_4/sdq3_5	0.332	0.061	0.000
sdq8_3/sdq8_4	0.282	0.129	0.028
sdq8_3/sdq8_5	−0.312	0.172	0.069
sdq8_4/sdq8_5	0.096	0.127	0.450
sdq13_3/sdq13_4	0.399	0.104	0.000
sdq13_3/sdq13_5	0.219	0.125	0.081
sdq13_4/sdq13_5	0.179	0.107	0.095
sdq16_3/sdq16_4	0.181	0.097	0.062
sdq16_3/sdq16_5	0.050	0.114	0.659
sdq16_4/sdq16_5	0.420	0.086	0.000
sdq24_3/sdq24_4	0.060	0.200	0.763
sdq24_3/sdq24_5	−0.019	0.265	0.943
sdq24_4/sdq24_5	0.034	0.203	0.865
Scaling factors	**Standardised coefficient**		
sdq5_3	0.483		
sdq7_3	0.710		
sdq12_3	0.466		
sdq18_3	0.581		
sdq22_3	0.698		
sdq5_4	0.457		
sdq7_4	0.684		
sdq12_4	0.440		
sdq18_4	0.553		
sdq22_4	0.672		

TABLE 39 Full model (hypothesis 1): measurement part 2 (*continued*)

Scaling factors	Standardised coefficient
sdq5_5	0.448
sdq7_5	0.675
sdq12_5	0.431
sdq18_5	0.543
sdq22_5	0.662
sdq3_3	0.767
sdq8_3	0.417
sdq13_3	0.531
sdq16_3	0.550
sdq24_3	0.376
sdq3_4	0.764
sdq8_4	0.415
sdq13_4	0.528
sdq16_4	0.547
sdq24_4	0.374
sdq3_5	0.766
sdq8_5	0.417
sdq13_5	0.530
sdq16_5	0.549
sdq24_5	0.375

TABLE 40 Full model (hypothesis 1): structural part 1

Cross-lagged effects of within-person factors	Coefficient	SE	*p*-value	Standardised coefficient
Conduct problems T4 regressed on				
Conduct problems T3	0.557	0.080	0.000	0.486
Emotional symptoms T3	−0.195	0.082	0.017	−0.170
Reading attainment T3	0.063	0.042	0.136	0.055
Emotional symptoms T4 regressed on				
Conduct problems T3	−0.152	0.137	0.269	−0.150
Emotional symptoms T3	−0.049	0.066	0.459	−0.049
Reading attainment T3	0.001	0.029	0.981	0.001
Reading attainment T4 regressed on				
Conduct problems T3	0.005	0.052	0.922	0.005
Emotional symptoms T3	0.010	0.025	0.692	0.010
Reading attainment T3	−0.113	0.011	0.000	−0.112
Conduct problems T5 regressed on				
Conduct problems T4	0.605	0.075	0.000	0.579
Emotional symptoms T4	−0.288	0.071	0.000	−0.244
Reading attainment T4	0.028	0.025	0.264	0.024

TABLE 40 Full model (hypothesis 1): structural part 1 (*continued*)

Cross-lagged effects of within-person factors	Coefficient	SE	*p*-value	Standardised coefficient
Emotional symptoms T5 regressed on				
Conduct problems T4	−0.068	0.120	0.569	−0.078
Emotional symptoms T4	0.038	0.086	0.655	0.039
Reading attainment T4	−0.005	0.021	0.812	−0.005
Reading attainment T5 regressed on				
Conduct problems T4	−0.014	0.067	0.836	−0.016
Emotional symptoms T4	−0.005	0.051	0.918	−0.005
Reading attainment T4	−0.151	0.011	0.000	−0.151
Covariance within wave of within-person factors				
Conduct problems T3/emotional symptoms T3	0.186	0.124	0.135	
Conduct problems T3/reading attainment T3	0.028	0.045	0.543	
Emotional symptoms T3/reading attainment T3	0.017	0.022	0.451	
Conduct problems T4/emotional symptoms T4	0.467	0.079	0.000	
Conduct problems T4/reading attainment T4	0.061	0.042	0.153	
Emotional symptoms T4/reading attainment T4	0.008	0.031	0.799	
Conduct problems T5/emotional symptoms T5	0.421	0.060	0.000	
Conduct problems T5/reading attainment T5	0.040	0.038	0.291	
Emotional symptoms T5/reading attainment T5	−0.070	0.029	0.017	
Covariance between latent random intercepts (between person)				
Conduct problems/emotional symptoms	0.413	0.124	0.001	
Conduct problems/reading attainment	−0.223	0.051	0.000	
Emotional symptoms/reading attainment	−0.118	0.023	0.000	

TABLE 41 Full model (hypothesis 1): structural part 2

Time invariant variables	Coefficient	SE	*p*-value	Standardised coefficient
Conduct problems regressed on				
Trial (if GBG)	0.015	0.068	0.824	0.010
Sex (if male)	0.846	0.068	0.000	0.549
Shared risk	0.643	0.052	0.000	0.263
Emotional symptoms regressed on				
Trial (if GBG)	0.017	0.055	0.749	0.012
Sex (if male)	−0.252	0.055	0.000	−0.173
Shared risk	0.525	0.043	0.000	0.227
Reading attainment regressed on				
Trial (if GBG)	−0.014	0.027	0.609	−0.009
Sex (if male)	−0.168	0.027	0.000	−0.115
Shared risk	−0.571	0.021	0.000	−0.246

TABLE 41 Full model (hypothesis 1): structural part 2 (*continued*)

Time invariant variables	Coefficient	SE	*p*-value	Standardised coefficient
Variance/covariance between observed variables				
Trial (if GBG) (variance)	0.250			
Trial (if GBG)/sex (if male)	−0.012			
Trial (if GBG)/shared risk	0.024			
Sex (if male) (variance)	0.249			
Sex (if male)/shared risk	0.032			
Shared risk (variance)	0.397			
Means of observed variables				
Reading attainment T3	0.317	0.029	0.000	
Reading attainment T4	0.318	0.029	0.000	
Reading attainment T5	0.385	0.032	0.000	
Trial (if GBG)	0.507	Exogenous		
Sex (if male)	0.525	Exogenous		
Shared risk	0.453	Exogenous		

TABLE 42 Model fit comparison hypothesis 0–hypothesis 1

Hypothesis	Degrees of freedom	χ-squared	χ-squared difference	Degrees of freedom difference	*p*-value
Hypothesis 1	587	19625.42			
Hypothesis 0	596	23608.54	954.831	9	0.000