# Good Behaviour Game

## Manchester Institute of Education, University of Manchester

| Evaluation Summary | |
|---|---|
| **Age range** | 7-10 |
| **Number of pupils** | 2,880 |
| **Number of schools** | 74 |
| **Design** | Efficacy |
| **Primary Outcome** | Behaviour |

## 1. Intervention

The Good Behaviour Game (hereafter referred to as GBG) is one of the most popular behaviour management systems for primary-aged children. It has an extensive evidence base supporting its use. Since its initial development over 40 years ago (Barrish, Saunders, & Wolf, 1969) multiple trials across the United States, the Netherlands and Belgium have attested to its effectiveness in promoting a range of positive outcomes (e.g. increased pro-social behaviour, reduced substance abuse, aggression and criminality) (Kellam et al., 2011). This has led to a number of endorsements from agencies such as the Substance Abuse and Mental Health Services Administration (who have included it on their National Registry of Evidence-Based Programmes and Practices) and the University of Colorado Blueprints For Healthy Youth Development (who have classified it as a ‘promising programme’ in their database).

The GBG itself can be described as an, “interdependent group-oriented contingency management procedure” (Tingstrom, Sterling-Turner, & Wilczynski, 2006, p. 225). Pupils in a given class are divided into mixed teams with up to 7 members. Strata can include a range of factors such as behaviour, academic ability, and gender. The teams then attempt to win the game as a means to access particular privileges/rewards. During the game period, the class teacher records the number of infractions to the following four rules among the teams: *(1) We will work quietly, (2) We will be polite to others, (3) We will get out of seats with permission, and (4) We will follow directions*. The game is ‘won’ by the team with the lowest number of infractions when it ends, although any team with fewer than four infractions also accesses the agreed reward (Chan, Foxcroft, Smurthwaite, Coombes, & Allen, 2012; Kellam et al., 2011).

Over the course of implementation of the GBG, there is a natural evolution in terms of the types of rewards used (from tangible rewards such as stickers to more abstract rewards such as free time), how long the game is played for (from 10 minutes to a whole lesson), at what frequency (from three times a week to every day), and when rewards are given (at the end of the game, end of the day, and at end of the week) (Elswick & Casey, 2011; Tingstrom et al., 2006). At face value, the GBG draws upon the principles of behaviour management – children receive reinforcement when they engage in appropriate behaviours. However, the group-orientation means that the intervention also uses principles of social learning theory – pupils at-risk of developing behaviour problems are able to learn from the appropriate behaviour being modelled by other team members. Finally, the GBG is informed by social field theory, which posits that successful adaptation at different life stages is contingent upon an individual’s ability to meet particular social task demands. In school, these task demands include being able to pay attention, work well with others, and obey rules. Success in social adaptation is rated both formally and informally by other members of the social field (e.g. teachers, peers). Social field theory predicts that improving the way in which teachers socialise children will improve their social adaptation. It is also predicted that early improvements in social adaptation will lead to better adaptation to other social fields later in life (Kellam et al., 2011).

The current education policy context is one that appears to be ideal for trialling approaches like the GBG. The White Paper *The Importance of Teaching* (Department of Education, 2011) set out a number of concerns about poor behaviour and its management in schools, and the recent changes to the Ofsted framework have made behaviour a more prominent feature of school inspections (see http://www.ofsted.gov.uk/resources/framework-for-school-inspection-january-2012).

## 2. Design

A 2-year cluster-randomised trial will be utilized. Participating schools will be the unit of randomisation. The allocation procedure will be conducted independently by the Manchester Academic Health Science Centre Clinical Trials Unit.

A minimization algorithm will be applied to the randomization to ensure balance across the arms of the trial in terms of the proportion of children eligible for free school meals (FSM). Schools allocated to the intervention arm of the trial will be trained and supported to implement the GBG over a two-year period. Schools allocated to the control arm of the trial will continue practice as usual during the same period. An incentive of £1500 will be offered to control schools - £1000 at the beginning of the trial and £500 following the completion of post-test measures[1]. Schools in both arms of the trial will be required to sign a memorandum of agreement committing them to comply with the evaluation protocol (defined as at least 85% completion rate of implementation and outcome measures).

### 2.1 Hypotheses

H1 - Children in primary schools implementing the GBG over a two-year period will demonstrate measurable improvements in reading (1a) and behaviour (1b) when compared to those children attending control schools

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

H3 – The effects outlined in H1 above will be amplified for children eligible for FSM

H4 – The effects outlined in H1 above will be moderated by variation in implementation

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

### 2.2 Sample

*2.2.1 Schools*

72 primary schools will be recruited from Local Authorities across the North West, South Yorkshire and East Midlands regions[2]. This figure represents the number of schools required to power the trial (66 - see below) plus an additional 6 to allow for attrition. Schools must be state-maintained and not already implementing the GBG in order to be eligible to participate.

[1] The only restriction on the use of this money will be that schools are not allowed to spend it on the Good Behaviour Game.
[2] These are regions in which the evaluation and implementation teams have strong networks.

### *2.2.2 Pupils*

Pupils in Year 3 (aged 7-8) at the beginning of the 2015/16 school year are the target cohort. We estimate a sample of 2,880 – an average of 40 per school – based on the proportion of single/mixed, double and triple form entry schools recruited in previous trials (e.g. PATHS to Success).

Baseline assessment will allow us to identify a subsample of boys at-risk of or already experiencing conduct problems for whom we predict an amplified effect of the GBG relative to the sample as a whole (following Kellam, Rebok, Ialongo, & Mayer, 1994). Based on national norms for our screening measure (www.sdqinfo.org/norms/UKNorm11.pdf; see Assessment of Outcomes below) we anticipate that 259 (18%) of the 1,440 boys in the sample will be identified as at-risk.

A second subsample – children eligible for FSM – will also be analysed in view of EEF's primary remit of narrowing the attainment gap for such students and in line with differential gains established for children from poorer socio-economic backgrounds for related universal programmes (e.g. Holsen, Iversen, & Smith, 2009). Experience from previous trials tells us to expect a higher proportion of such pupils than the national average (19.2% - Department for Education, 2013) – we estimate c.30% (N=864).

### *2.2.3 Techers*

We estimate approximately 107 teachers (2880 divided by an average class size of c.27 - Department for Education, 2011) across arms in each year of the trial (Year 3 in first year, Year 4 in second year), providing a sample of 214. This figure does not account for mixed year group teaching and other potential complicating factors (including increased attrition/noncompliance in teacher outcome measures), so a revised estimate of 150 will be used. Without adjusting for the design effect (e.g. clustering of teachers within schools), this will allow us to detect an effect size of around 0.45 for our teacher-level secondary outcome measures with Power at 0.8 and Alpha at 0.05.

### *2.2.4 Power and sample size calculations*

Using a demographic and pre-test covariate model, we assume an intra-cluster correlation co-efficient (ICC) of no more than 0.06 for our primary outcome measure (Hedges & Hedberg, 2007). Given this, and standard Power and Alpha thresholds of 0.8 and 0.05 respectively, the above sample of schools and pupils will be sufficient to detect an effect size of 0.2 in an intention-to-treat analysis. This will also allow the detection of an effect size of 0.3 or higher in the at-risk subsample.

## 2.3 ASSESSMENT OF OUTCOMES

The primary outcome measure for this study is children's attainment in *reading*[3]. Secondary outcome measures are children's *behaviour*, teacher *efficacy in classroom management, classroom stress*, and *retention*. The primary outcome will be assessed at pre-test (T1) and post-test (T3) only. The secondary outcome measures will be assessed at pre-test (T1), after 12 months (T2), and at post-test (T3).

[3] Reading has been selected for two reasons. First, to align the trial with previous studies of the GBG (e.g. Dion et al, 2011). Second, improvements in reading are a theoretically plausible distal consequence of proximal changes in attention brought about by the intervention. Third, assessment of both reading *and* maths would yield a significant data burden on schools.

The primary outcome measure at T1 will be derived from the National Pupil Database (see below). At T3 it will be administered via on-site whole-class testing. All other outcome measures will be administered online using World App Key Survey.

*2.3.1 Reading*

The baseline period for the trial coincides with the end of Key Stage 1 teacher assessments for the study cohort and so children's KS1 National Curriculum reading levels will be used as the pre-test covariate.

Post-test assessment of reading will utilize the Hodder Group Reading Test (www.hoddertests.co.uk). This paper-based measure produces National Curriculum levels, reading ages and standardized scores. It can be administered in a whole-class/group context and takes 30-35 minutes to complete, minimizing the data burden for participating schools.

Assessment of the primary outcome measure will employ a single-blind approach – the research staff administering the test will not be aware of the allocation status of individual schools.

*2.3.2 Behaviour*

Children's behaviour will be assessed using the Teacher Observation of Children's Adaptation checklist (TOCA-C; Koth, Bradshaw, & Leaf, 2009). This 21-item scale provides indices of children's concentration, disruptive behaviour and pro-social behaviour. Raters read statements about a child (e.g. "Pays attention") and endorse them on a 6-point scale (Never/Rarely/Sometimes/Often/Very Often/Almost Always).

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

*2.3.4 Teacher efficacy in classroom management*

Teacher efficacy in classroom management will be assessed using the 4-item subscale of the short-form Teachers' Sense of Efficacy Scale (TES; Tschannen-Moran & Hoy, 2001). Teachers read questions (e.g. "How much can you control disruptive behaviour in the classroom?") and respond on a 9-point scale (from Not At All to A Great Deal).

*2.3.5 Teacher classroom stress*

Teacher stress will be captured using the 3-item classroom stress subscale of the Teacher Stress Inventory (TSI; Boyle, Borg, Falzon, & Baglioni, 1995). Respondents read questions (e.g. "How great a source of stress is maintaining class discipline?") and respond on a 5-point scale (from No Stress to Extreme Stress).

*2.3.6 Teacher retention*

Teacher retention will be assessed through the use of a single item measure, as follows: "How likely are you to leave the teaching profession in the next 5 years?" Participating teachers will respond on a 6-item scale (Definitely/Highly Likely/Likely/Unlikely/Highly Unlikely/Definitely Not)

## 2.4 Assessment of implementation

In our assessment of implementation we will seek to determine:

1. Fidelity/integrity – to what extent do teachers adhere to the GBG guidance?
2. Dosage/exposure – how frequently is the GBG played and for how long?
3. Quality – how well do teachers deliver the components of the GBG?
4. Participant responsiveness – to what extent do children engage with the GBG?
5. Reach – what is the rate and scope of participation in the GBG across the class?
6. Programme differentiation – to what extent can the GBG be distinguished from other, existing behaviour management practices in both intervention and control schools?
7. Adaptations – what is the nature and extent of changes made to the GBG during the course of implementation?

Our data collection protocol will be informed by those used in previous GBG studies (e.g. Elswick & Casey, 2011), our own work in other trials (e.g. PATHS to Success), and naturally occurring data produced through the course of the intervention (e.g. the 'game' and 'probe' data relating to rule infraction could be used as a proxy for participant responsiveness - Chan, Foxcroft, Smurthwaite, Coombes, & Allen, 2012). A critical consideration in relation to programme differentiation is what constitutes usual/existing practice in intervention (to allow us to identify foundational practices that may interact with the GBG) and control schools (to allow us to establish the counter-factual). To this end we will conduct a school-level survey across all participating schools ahead of each wave of outcome data collection that explores usual/existing behaviour management and related practices at both the universal and targeted levels.

For the other aspects of implementation, two primary options are available from a methodological standpoint – teacher self-report and independent observation.

### *2.4.1 Teacher self-report*

We will survey the teachers implementing the GBG with our study cohort via World App Key Survey in the second half of the spring term (e.g. late February to April). The survey will comprise a series of structured items relating to implementation factors noted above, to which teachers will provide a response on a fixed scale (e.g. "Please rate the extent to which, on average, you have adhered to the GBG guidance in your implementation during the current school year", rated from 1 Not At All to 10 Completely). Each subsection will also include an open-field text box in which teachers will be invited to note any factors they feel have influenced that aspect of their implementation.

This approach is the most cost-effective and enables teachers to reflect and report upon a sustained period of implementation activity. However, data produced is subject to demand characteristics and impression management (Humphrey, 2013). It is a less reliable and valid estimate than independent observations.

### *2.4.2 Independent observations*

We will develop a structured observation schedule that will be piloted and refined ahead of the project's empirical phase in order to establish inter-rater reliability[4] and ensure that the

[4] Average Cohen's Kappa of 0.7 or above will be required.

measure is fit for purpose[5]. The schedule will comprise a series of structured items relating to the implementation factors noted above, to which observers will provide a response on a fixed scale (e.g. “Rate the extent to which the teacher plays the GBG with interest and enthusiasm”, from 1 Not At All to 10 Completely). An annotated scoring rubric with prompts for each aspect of implementation derived from the pilot will accompany the schedule. Open-field text boxes will be included to allow observers to record reasons for their ratings.

One observation per teacher/class will be conducted in each year of the trial during the spring term (e.g. January to April). 10% of observations will be moderated by a senior member of the research team to ensure that inter-rater reliability established *a priori* is maintained throughout the trial.

This approach is more costly than implementer self-report but benefits from greatly increased rigour and validity. Data from independent observations are much more closely aligned with intervention outcomes (Domitrovich, Gest, Jones, Gill, & DeRousie, 2010), with implementer self-reports not being significantly associated with outcome measures in several studies (e.g., Goldberg Lillehoj, Griffin, & Spoth, 2004; Social and Character Development Research Consortium, 2010). In consequence, we strongly recommend the use of independent observations in preference to implementer surveys.

## 2.5 Assessment of process

We will conduct longitudinal case studies of up to 6 GBG schools purposively sampled using a maximum variation approach (e.g. existing behavior management practices, perceived levels of behaviour/discipline problems, Ofsted behaviour and safety ratings). The case studies will explore issues of social validity, acceptability and feasibility of the GBG (e.g. does the intervention meet schools’ perceived needs? How well received is the intervention among staff and pupils? Can the intervention be delivered successfully?). Here we will draw upon relevant studies of school-based interventions (e.g. Kendal, Callery, & Keeley, 2011) and adapt existing rubrics from the implementation literature (e.g. Bird et al., 2014) to inform our data generation. In addition to this, we will use the case studies to explore a range of factors affecting implementation at the different domains/levels consistently identified in the literature: preplanning and foundations (e.g. buy-in), implementation support system (e.g. ongoing external support), implementation environment (e.g. time constraints), implementer factors (e.g. experiences, skills and confidence in delivery), and programme characteristics (e.g. flexibility) (Durlak & DuPre, 2008; Forman, Olin, Hoagwood, & Crowe, 2009; Greenberg, Domotrovich, Graczyk, & Zins, 2005). We anticipate the use of interviews, focus groups, observations and document analysis, drawing upon the views of a range of informants (e.g., pupils, teachers, school leaders, parents, GBG coaches). At least one two-day fieldwork visit per term per school over the two years of the trial will likely be required.

## 2.6 Analytical strategy

### *2.6.1 Quantitative data*

For H1 we will conduct intention-to-treat analysis (Gupta, 2011), operationalized as three-level hierarchical linear models (time, child, school), with intervention group (e.g. GBG vs. control) and minimization variables (e.g. % FSM) entered at the school level. For H2 and H3 these will be extended to include risk-status at the child level (e.g. normal vs. borderline/abnormal) such that 3-way cross-level interactions (intervention group*risk status*time) can be examined. For H4 we will conduct adherence to protocol/per protocol

[5] This will be achieved either through live observations of GBG implementation in existing sites (e.g. Oxford) or using video footage if this is available from the programme developers.

analyses using three-level hierarchical linear models (child, class/teacher, school) with data from GBG schools only, with implementation data (see 2.4) entered at the class/teacher and school levels. For H5 we will conduct multivariate analyses of co-variance (MANCOVA) to reduce the family-wise error rate in view of the restricted sample. In all cases, effect sizes will be reported using Hedge's g (Cohen's d bias corrected) and accompanied by 95% confidence intervals as per EEF specifications.

### *2.6.2 Qualitative data*

Qualitative data will be treated in two ways. First, we will produce detailed case profiles of each school that document their first two years of implementation, paying attention to how individual context and circumstances have influenced progress in each. Second, we will thematically analyse our data corpus using the principles and processes outlined by Braun and Clarke (2006) (e.g. familiarization, generating initial codes, searching for themes, reviewing themes, defining and naming themes, report production). A hybrid approach will be taken which will likely be informed by conceptual models of implementation in school settings (e.g. Domitrovich et al., 2008), our primary orienting concepts (social validity, acceptability, feasibility) and the GBG implementation rubrics, while also allowing for the emergence of unanticipated themes specific to this project/context.

## 3. Project timescale

See Gantt chart below.

# Good Behaviour Game

## Manchester Institute of Education, University of Manchester

| Activities | Lead-in | 1<br>09/14 | 2 | 3 | 4 | 5<br>2015 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17<br>2016 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29<br>2017 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Project management** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Ethics approval, contracts, set-up | 🟩 | 🟩 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| EEF/Mentor/UoM meetings | 🟩 | | | | | | | | | | | | | 🟩 | | | | | | | | | | | | 🟩 | | | | | | | | | | | | 🟩 | |
| School recruitment | | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| NPD extraction | | | | | | 🟩 | 🟩 | 🟩 | 🟩 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Develop project website & update | | | | | | 🟩 | 🟩 | 🟩 | 🟩 | | | | | | | | | | | 🟩 | 🟩 | | | | | | | | | | | 🟩 | 🟩 | | | | | | |
| Randomisation | | | | | | | | | | | 🟩 | 🟩 | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| GBG training | | | | | | | | | | | | | | 🟩 | | | | | | | | | | | | 🟩 | | | | | | | | | | | | | |
| School coaching | | | | | | | | | | | | | | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | | | |
| EEF report writing | | | | | | | 🟩 | | | | | | 🟩 | | | | | | 🟩 | | | | | | 🟩 | | | | | | 🟩 | | | | | | 🟩 | 🟩 | 🟩 |
| **Assessment of outcomes** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Instrumentation | 🟧 | 🟧 | 🟧 | 🟧 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Outcome measures (child) | | | | | | | | | 🟧 | 🟧 | 🟧 | | | | | | | | | | | 🟧 | 🟧 | | | | | | | | | | | 🟧 | 🟧 | | | | |
| Outcome measures (teacher) | | | | | | | | | | | | | | 🟧 | | | | | | | | | 🟧 | | | 🟧 | | | | | | | | | 🟧 | | | | |
| Data cleaning and analysis | | | | | | | | | | | | 🟧 | 🟧 | 🟧 | 🟧 | | | | | | | | | 🟧 | 🟧 | 🟧 | 🟧 | | | | | | | | | 🟧 | 🟧 | 🟧 | 🟧 |
| **Assessment of implementation** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Develop & pilot measures | 🟪 | 🟪 | 🟪 | 🟪 | 🟪 | 🟪 | 🟪 | 🟪 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| School surveys | | | | | | | | | 🟪 | | | | | | | | | | | | 🟪 | | | | | | | | | | | | 🟪 | | | | | | |
| Classroom | | | | | | | | | | | | | | | | | | 🟪 | 🟪 | 🟪 | | | | | | | | | | 🟪 | 🟪 | 🟪 | | | | | | | |

# Good Behaviour Game

## Manchester Institute of Education, University of Manchester

| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| observations | | | | | | | | | | | | | | | | | | 🟪 | 🟪 | 🟪 | | | | | | | | | | 🟪 | 🟪 | 🟪 | | | | | | | |
| Teacher surveys | | | | | | | | | | | | | | | | | | | 🟪 | 🟪 | | | | | | | | | | | 🟪 | 🟪 | | | | | | | |
| Data cleaning and analysis | | | | | | | | | | 🟪 | | | | | | | | | | 🟪 | 🟪 | 🟪 | | | | | | | | | | 🟪 | 🟪 | 🟪 | | | | | |
| **Assessment of process** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Develop & pilot data generation tools | 🟦 | 🟦 | 🟦 | 🟦 | 🟦 | 🟦 | 🟦 | 🟦 | 🟦 | 🟦 | 🟦 | 🟦 | 🟦 | 🟦 | | | | | | | | | | | | | | | | | | | | | | | | | |
| Fieldwork visits | | | | | | | | | | | | | | | 🟦 | | | | 🟦 | | | 🟦 | | | | | 🟦 | | | | 🟦 | | | 🟦 | | | | | |
| Transcription | | | | | | | | | | | | | | | | 🟦 | | | | 🟦 | | | 🟦 | | | | | 🟦 | | | | 🟦 | | | 🟦 | | | | |
| Data cleaning and analysis | | | | | | | | | | | | | | | | | 🟦 | | | | 🟦 | | | 🟦 | | | | | 🟦 | | | | 🟦 | | | 🟦 | 🟦 | 🟦 | 🟦 |

**Table 1. Gantt chart for Good Behaviour Game Trial (proposed start 09/14; proposed end 10/17).**

### 4. References

Barrish, H. H., Saunders, M., & Wolf, M. M. (1969). Good behavior game: effects of individual contingencies for group consequences on disruptive behavior in a classroom. *Journal of Applied Behavior Analysis*, *2*, 119–24.

Bird, V. J., Le Boutillier, C., Leamy, M., Williams, J., Bradstreet, S., & Slade, M. (2014). Evaluating the feasibility of complex interventions in mental health services: standardised measure and reporting guidelines. *The British Journal of Psychiatry : The Journal of Mental Science*, *204*, 316–21.

Boyle, G. J., Borg, M. G., Falzon, J. M., & Baglioni, A. J. (1995). A structural model of the dimensions of teacher stress. *The British Journal of Educational Psychology*, *65*, 49–67.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*, 77–101.

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

Department for Education (2011). *Class size and education in England: evidence report*. London: DFE.

Department for Education (2013). *Schools, pupils and their characteristics: January 2013*. London: DFE.

Department of Education (2011). *The importance of teaching*. London: DFE.

Domitrovich, C. E., Bradhsaw, C. P., Poduska, J. M., Hoagwood, K., Buckley, J. A., Olin, S., … Ialongo, N. S. (2008). Maximising the implementation quality of evidence-based preventive interventions in schools: a conceptual framework. *Advances in School Mental Health Promotion*, *1*, 6–28.

Domitrovich, C. E., Gest, S. D., Jones, D., Gill, S., & DeRousie, R. M. S. (2010). Implementation quality: Lessons learned in the context of the Head Start REDI trial. *Early Childhood Research Quarterly*, *25*, 284–298.

Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: a review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology*, *41*, 327–50.

Elswick, S., & Casey, L. (2011). The good behavior game is no longer just an effective intervention for students: An examination of the reciprocal effects on teacher behaviors. *Beyond Behavior*, *21*, 36–46.

Forman, S., Olin, S., Hoagwood, K., & Crowe, M. (2009). Evidence-based interventions in schools: developers' views of implementation barriers and facilitators. *School Mental Health*, *1*, 26–36.

Goldberg Lillehoj, C. J., Griffin, K. W., & Spoth, R. (2004). Program provider and observer ratings of school-based preventive intervention implementation: agreement and relation to youth outcomes. *Health Education & Behavior*, *31*, 242–57.

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. *Journal of Child Psychology and Psychiatry*, *38*, 581–586.

Greenberg, M., Domotrovich, C., Graczyk, P., & Zins, J. (2005). *The study of implementation in school-based preventive interventions: theory, research and practice*. Washington: USDHHS.

Gupta, S. K. (2011). Intention-to-treat concept: A review. *Perspectives in Clinical Research*, *2*, 109–12.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*, 60–87.

Holsen, I., Iversen, A. C., & Smith, B. H. (2009). Universal social competence promotion programme in school: Does it work for children with low socio-economic background? *Advances in School Mental Health Promotion*, *2*, 51–60.

Humphrey, N. (2013). *Social and emotional learning: a critical appraisal*. London: Sage Publications.

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

Kellam, S. G., Rebok, G. W., Ialongo, N., & Mayer, L. S. (1994). The course and malleability of aggressive behavior from early first grade into middle school: results of a developmental epidemiologically-based preventive trial. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *35*, 259–81.

Kendal, S., Callery, P., & Keeley, P. (2011). The feasibility and acceptability of an approach to emotional wellbeing support for high school students. *Child and Adolescent Mental Health*, *16*, 193–200.

Koth, C. W., Bradshaw, C. P., & Leaf, P. J. (2009). Teacher Observation of Classroom Adaptation--Checklist: Development and factor structure. *Measurement and Evaluation in Counseling and Development*, *42*, 15–30.

Social and Character Development Research Consortium. (2010). *Efficacy of school-wide programs to promote social and character development and reduce problem behavior in elementary school children*. Washington, DC: SCDRC.

Tingstrom, D. H., Sterling-Turner, H. E., & Wilczynski, S. M. (2006). The good behavior game: 1969-2002. *Behavior Modification*, *30*, 225–53.

Tschannen-Moran, M., & Hoy, A. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and Teacher Education*, *17*, 783–805.