Abstract

This mixed-methods study examines two moderators of the impact of the Good Behavior Game – implementation variability, participant risk status, and the interaction between them – as predictors of behavioral and academic outcomes. Quantitative data from 38 primary schools were utilized, with outcome data collected at baseline and two-year follow-up. Behavior (disruptive behavior, pro-social behavior, and concentration problems) was assessed via the Teacher Observation of Classroom Adaptation Checklist. Reading attainment was assessed via national teacher assessment scores, and the Hodder Group Reading Test. Implementation fidelity/quality data were collected via independent observations. Participant risk status was modeled using a cumulative risk index. Multi-level modeling revealed that higher levels of fidelity/quality were associated with improved overall reading scores (d=.203-225), but worsening disruptive behavior among high-risk students (d=.560). Thematic analysis of qualitative interview data collected from 20 teachers identified six groups of at-risk students who were perceived to experience differential effects, and five key mechanisms underpinning these.

Beyond 'what works': A mixed-methods study of intervention effect modifiers in the Good Behavior Game

This paper focuses on the Good Behavior Game (GBG), a universal, evidence-based interdependent group-contingency behavior management strategy (Barrish, Saunders, & Wolf, 1969; Lastrapes, 2013). We use data from the intervention arm of our recent randomized trial of the GBG to advance knowledge in relation to two key moderators of the impact of universal, school-based interventions: implementation variability and participant risk status, and the interaction between them. Our focus reflects a recent paradigm shift in the evaluation of school-based interventions, with researchers moving beyond the question of 'what works' to examine "what works for whom, and under what circumstances" (Bonell, Fletcher, Morton, Lorenc, & Moore, 2012, p. 2303). To this end, some have focused on *how* and *why* interventions work by examining how implementation variability moderates outcomes (e.g. Humphrey, Barlow, & Lendrum, 2017), while others have focused on *who* interventions work for by examining differential effects for specific population subgroups (e.g. Dolan et al., 1993; Kellam et al., 2011). However, to date, there has been a paucity of studies at the intersection of these two important areas of inquiry (i.e., those that have explored whether implementation matters more for certain groups). The current study was designed to address this critical gap in the knowledge base. In the following sections we briefly review implementation and subgroup research respectively, before returning to our rationale regarding their intersection.

Implementation Matters

A growing body of evidence suggests that the way school-based interventions are implemented can impact their success (Durlak, 2016). Implementation is "the process by which an intervention is put into place" (Lendrum & Humphrey, 2012, p. 635), the most commonly measured aspects of which are *fidelity* (the extent to which key aspects are

delivered as intended), *quality* (how well different aspects are delivered), and *dosage* (how much of the intended program is delivered; Durlak & DuPre, 2008). The study of implementation arose in part due to concerns about the "black box" approach to evaluation, which focuses on whether or not an intervention worked, without looking at what actually happened in order to establish how or why outcomes were affected (Harachi, Abbott, Catalano, Haggerty, & Fleming, 1999). Implementation and process evaluations thus increase internal validity and protect against Type III errors (the inaccurate attribution of cause; Lendrum & Humphrey, 2012). For example, in the context of null results, it helps to determine whether this was due to poor program design, or to poor implementation of a well-designed program (Askell-Williams, Dix, Lawson, & Slee, 2012). Confirming that the key program components and processes have been implemented means that links can be made between the achieved outcomes and the intervention (Lendrum & Humphrey, 2012).

With the exception of social and emotional learning interventions (Wigelsworth et al., 2016), routine reporting and analysis of implementation data in studies of school-based preventive interventions is still relatively uncommon (Bruhn, Hirsch, & Lloyd, 2015; Hagermoser Sanetti, Dobey, & Gallucci, 2014). Those that have explored the relationship between implementation variability and student outcomes have typically found that greater outcomes are achieved when the intervention is implemented as intended (O'Donnell, 2008). For instance, a recent English study of the Promoting Alternative Thinking Strategies curriculum (PATHS) found that both higher implementation quality and participant responsiveness were associated with significantly lower ratings of externalizing problems (Humphrey et al., 2017). Similarly, an evaluation of the Australian KidsMatter mental health initiative found that social and emotional competences improved significantly more in average- and high-implementing schools (compared to low-implementing schools; Askell-Williams et al., 2012). However, studies examining the effects of implementation variability

of the GBG are rare. Indeed, Donaldson, Vollmer, Krous, Downs, & Berard (2011) recommended "one area for future research could involve systematically evaluating the effects of changes in treatment integrity on the effectiveness of the GBG" (p.607). While a handful of studies measure and report levels of GBG fidelity and dosage (e.g., Domitrovich et al., 2015; Hagermoser Sanetti & Fallon, 2011), these data have rarely been used to establish potential moderating associations with intervention outcomes. One notable exception is a study conducted by Ialongo et al. (1999), which we discuss in more detail below.

Universal Intervention: Differential Gains?

Notwithstanding the moderating influence of implementation, there is also evidence to suggest that outcome variability may be driven in part by participant characteristics. Natural heterogeneity exists within universal populations, and universal interventions can differentially affect various strata of the population (Greenberg & Abenavoli, 2017); indeed, universal, school-based interventions seem to be particularly beneficial for certain at-risk subgroups (Jones, Brown, & Aber, 2011). For example, an evaluation of Second Step, a universal preventive intervention, found some evidence of differential gains among students from socio-economically disadvantaged backgrounds in terms of social competence, school performance and life satisfaction (Holsen, Iversen, & Smith, 2009).

In relation to the GBG, the intervention has been found to yield particularly beneficial behavioral outcomes among highly aggressive males during middle childhood. These effects appear to be long-lasting, with this group of individuals demonstrating reductions in drug abuse and dependence disorders, antisocial personality disorders, and incarceration for violence when followed up in early adulthood (Dolan et al., 1993; Kellam et al., 2011).

However, existing research has typically examined risk factors in isolation (e.g., males), when in reality they cluster and co-occur. In cumulative risk (CR) theory (Rutter,

1979) it is acknowledged that children are often exposed to *multiple* risk factors, with complex and interactional relationships between them (Gerard & Buehler, 1999). Indeed, it is the accumulation of risk factors that is theorized to lead to negative outcomes, with those at higher risk experiencing greater difficulties (Rutter, 1979). Only examining a single risk factor means that their apparent importance may be over-estimated (Sameroff, Gutman, & Peck, 2003). The adoption of a CR approach offers considerable promise in subgroup moderator analyses, as it more accurately represents individual differences and the multitude of factors influencing a child's development. To date, however, only one study has utilized a CR approach when examining the effectiveness of a school-based intervention (The Multisite Violence Prevention Project, 2008). They found that short- and long-term effects of the "Guiding Responsibility and Expectations in Adolescents Today and Tomorrow" (GREAT) student curriculum on social-cognitive factors varied as a function of students' pre-intervention level of risk. While high-risk students evidenced gains in self-efficacy and attitudes towards aggression and non-violent behavior, effects for low-risks students were in the opposite direction. Thus, while differential effects are often referred to as "gains", in reality, the impact of a given intervention on particular subgroups may be negative in some cases.

The Current Study

While there is evidence to suggest that an intervention does not affect all students equally, and the way that it is implemented can influence its success, the interaction between classroom-level implementation variability and student-level risk exposure in predicting intervention outcomes has not previously been systematically and rigorously examined. If groups of students respond differently to an intervention, then it follows that certain elements of said intervention are influencing these variable responses. It is theorized that universal interventions can produce several different types of outcomes, and that these outcomes can

vary for different subgroups, depending on the extent to which the individuals within them display deficiencies in the skills targeted by the intervention (Farrell, Henry, & Bettencourt, 2013; Greenberg & Abenavoli, 2017). As the GBG logic model outlines immediate improvements in disruptive behavior, consistent on-task behavior and increased pro-social behaviors (Chan, Foxcroft, Smurthwaite, Coomes, & Allen, 2012), it could be assumed that those most at-risk for difficulties in these areas would experience the greatest gains from strict adherence to the prescribed program procedures. However, the subgroup that will benefit the most is currently contested; while Muthén et al. (2002) posited that those exposed to moderate levels of risk would evidence the greatest gains from interventions such as the GBG, others have suggested the highest-risk participants would benefit more (Farrell et al., 2013; Greenberg & Abenavoli, 2017).

To date, extremely limited and tentative evidence exists for possible interactions between implementation variability and participant risk status in predicting intervention outcomes. One study of the GBG conducted by Ialongo et al. (1999) in Baltimore, USA, found that for males only, higher fidelity was associated with fewer nominations of aggressive behavior and higher reading achievement scores. However, there are several issues surrounding these findings that need to be addressed. First, while Ialongo's study did suggest that implementation variability may have differential effects for certain risk groups, the GBG was implemented alongside another intervention, creating a significant confound. Second, it is unclear *why* males responded differently at varying levels of implementation. Finally, as the authors only examined differential effects based on exposure to a single risk factor, it is not yet known whether implementation variability is associated with differential gains for children at varying levels of CR; this clearly warrants further attention.

In light of the above, the intended contribution of the current study is to improve understanding of the effects of the GBG by examining the interaction between levels of

implementation (fidelity/quality) and participant risk status (CR exposure) as predictors of behavioral (disruptive behavior, concentration problems, pro-social behavior) and academic (reading attainment) outcomes. We further extend research in this area by exploring *why* students at different levels of risk exposure may respond differently to varying levels of implementation. Thus, we also examine the perceived mechanisms underlying any differential effects. In asking not just 'what works', but what works for whom *and* under what conditions and circumstances, we aim to provide a unique and significant contribution to prevention and implementation science.

Methodology

Design

The current study utilizes a subset of quantitative and qualitative data collected as part of a two-year mixed-methods cluster-randomized controlled trial (RCT) of the GBG (Humphrey et al., 2018). Seventy-seven primary schools (N students = 3,084) in three regions across England were randomly allocated to one of two trial arms: (1) GBG (intervention arm; 38 schools); or (2) usual provision (UP arm; 39 schools). Teachers in schools allocated to the intervention arm were trained and supported to implement the GBG during the two-year trial period (2015/16 and 2016/17). The trial protocol is available here: https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/the-good-behaviour-game/. The data utilized in the current study were taken from the intervention arm of the trial; we thereby utilize a multi-level, natural variation design for the quantitative aspect of our work. Semi-structured teacher interview data drawn from longitudinal case studies of six self-selecting GBG schools in the trial's implementation and process evaluation (IPE) are also utilized. Collectively, this represents a sequential explanatory design, whereby qualitative data are used to illuminate and explain quantitative findings.

Participation required consent from the schools' Head Teachers. Child assent and parental opt-out consent were also sought. In total, 21 parents (1.35%) in the current study sample exercised their right to opt their children out of data collection, and no children declined assent or exercised their right to withdraw from the study. Opt-in consent was obtained from all teachers that participated in interviews. The study received approval from the ethics committee of the authors' host institution.

Participants

Quantitative strand. The target cohort were N=1560 children (aged seven-eight) in the 38 GBG schools noted above. The composition of these schools mirrored that of primary schools in England in relation to size and the proportion of students speaking English as an Additional Language (EAL), but contained significantly larger proportions of children with special educational needs and disabilities (SEND) and those eligible for free school meals (FSM), in addition to lower rates of absence and attainment. The student sample were also generally above the national average in terms of the proportion of children with an SEND, eligible for FSM, and speaking EAL; while they were generally below average with regards to attainment (DfE, 2015; Table 1).

Sixty-one teachers in the first year and 60 teachers in the second year (60 classes[1]) implemented the GBG. 80.2% were female, and teachers had been in the profession for an average of 8 years.

Qualitative strand. Twenty staff members (the 14 teachers who had delivered the intervention and 6 senior leadership team [SLT] members) were interviewed from the six case study schools. The schools were diverse in terms of their compositional and contextual characteristics (e.g., proportion of students eligible for FSM and identified as having an SEND). Eighty percent of these participants were female, and teachers had been in the

[1] There was an additional class-share in the first year

profession for an average of six years. Mean fidelity/quality scores among the case study schools were 66% (Table 1).

Intervention

Numerous versions and variations of the GBG have been developed, utilized and reported in the literature. In the current study, the American Institutes for Research version of the GBG was utilized (Ford, Keegan, Poduska, Kellam, & Littman, 2014). Core components of the GBG are (1) classroom rules, (2) team membership, (3) monitoring behavior, and (4) positive reinforcement. While playing the game, students in a class are divided into teams of up to seven. These are typically gender-balanced and heterogeneous in behavior and academic ability. Teams attempt to win the GBG in order to access certain rewards or privileges. To access these rewards, they need to have four or fewer infractions at the end of the game, which are recorded using "check-marks". At the beginning of the year, it is recommended that these rewards are tangible (e.g., stickers) and given immediately after the game ends; as the school year progresses, the rewards should become intangible (e.g., free time) and their receipt should be delayed (e.g., end of day/week).

While the game is being played, the teacher monitors behavior and records any infractions that occur as a result of a team member failing to follow one of four rules: (1) we will work quietly[2], (2) we will be polite to others, (3) we will get out of our seats with permission, and (4) we will follow directions (Kellam et al., 2011). After an infraction, teachers follow a "check-comment-redirect" script to identify the rule broken to the relevant team. Other than when recording an infraction, teachers should not directly interact with individual students during the game. It is recommended that initially the game be played three times a week, for ten minutes each time, increasing over the year to every day for up to 30 minutes. It should also be played at varying points throughout the day, during an

[2] Adherence to "quietly" is defined as working at a "voice level" set by the teacher that is deemed to be appropriate for a particular activity

assortment of lessons and activities. The game is designed to be integrated into the existing curriculum without taking up any additional teaching time.

Teachers in GBG schools attended two days of training prior to implementation, with a further day of top-up training later in the academic year. Trained coaches visited GBG schools approximately once per month throughout the course of the trial to support teachers' implementation efforts (e.g., through modeling of game sessions, observation and feedback; Ashworth, Demokwicz, Lendrum & Frearson., 2018).

Measures

Behavior. Teacher perceptions of behavior were assessed using the checklist version of the Teacher Observation of Classroom Adaptation (TOCA-C; Koth, Bradshaw & Leaf, 2009), which teachers completed for each pupil at baseline and the end of the trial, following two years of implementation.. This 21-item scale assesses students' concentration problems (inattentive and off-task behavior), disruptive behavior (disobedient, disruptive and aggressive behaviors) and pro-social behavior (positive social interactions). Statements are provided about the child (e.g., "gets angry when provoked by other children"), which teachers read and endorse on a 6-point scale (Never/Rarely/Sometimes/Often/Very Often/Almost Always). Responses are then summed for each subscale, with higher scores indicating more maladaptive behaviors for concentration and disruptive behavior, and lower scores indicative of less pro-social behavior (Kourkounasiou & Skordilis, 2014).

The TOCA-C has sound psychometric properties, including high internal consistency (all subscales $\alpha>0.86$), and has a factor structure that is invariant across gender, race and age (Bradshaw, Waasdorp, & Leaf, 2015; Koth et al., 2009). Internal consistency of the TOCA-C subscales in the current study was excellent (all $\alpha>0.85$ at baseline).

Reading. End of Key Stage 1 (KS1) national teacher assessment scores (specifically the KS1 National Curriculum reading point score: the KS1_READPOINTS variable) were

utilized as a pre-test measure of reading attainment. Reading data were extracted from the National Pupil Database (NPD) by the authors at baseline. These data are collected across England when children reach the end of Year 2 (age six-seven) and higher scores are indicative of greater reading attainment. KS1 scores are highly predictive of future academic performance, both in terms of Key Stage 2 (KS2) assessment scores (when children are 10-11; Humphrey et al., 2015) and independent standardized test scores (Humphrey et al., 2018).

The Hodder Group Reading Test (HGRT; specifically test sheet 2A) was utilized as a post-test measure of reading attainment. This measure was chosen as it produces scores that correlate as well with the Key Stage assessments noted above as said assessments do with each other (KS1-KS2 = .73; KS1-HGRT = .74; EEF, 2013; Humphrey et al., 2018). This means that the pre- and post-test variables were comparable without having to conduct additional baseline reading tests, thereby reducing data burden for schools.

The HGRT has been standardized on over 13,000 children ($\alpha = 0.95$; Devine, Soltész, Nobes, Goswami, & Szucs, 2013) and is capable of reliably measuring reading ability over a broad chronological age range between seven and 16 years. Higher scores are indicative of greater reading attainment. The research team administered the HGRT in a whole class context over a period of 30 minutes in the final term of the second year of the trial.

Fidelity/quality. Fidelity and quality were assessed via a structured observation schedule administered by a member of the research team once per year, in the spring term, between January and April. Regarding fidelity, a list of required steps outlined in the GBG manual (Ford et al., 2014) were scored on a binary yes/no scale. Quality was rated on a five-item scale of 0-2, with higher scores indicating higher quality of delivery. The observation schedule was developed for the purposes of the trial, and also incorporated items designed to measure participant responsiveness. It was piloted and refined using video footage of GBG implementation in English schools recorded in a UK pilot study (Chan et al., 2012). Inter-

rater reliability was tested and found to be “almost perfect” or above (intra-class coefficients [for ordinal items] >.74; Cohen’s Kappa [for nominal items] >.8; Hallgren, 2012).

Prior to analysis, an assessment of the structure of the schedule was conducted using exploratory factor analysis (EFA) with Weighted Least Squares Mean and Variance adjusted (WLSMV), while accounting for clustering in the data. Only items with factor loadings above .32 were retained (Tabachnick & Fidell, 2015). Parallel analysis indicated a two-factor structure for the observation schedule. Subsequently, a two-factor EFA was conducted, with all items loading substantially onto one of two domains: fidelity/quality and participant responsiveness. Thus, fidelity and quality were treated as one combined variable in analyses, with teachers receiving a percentage score between zero and 100. This was chosen as the measure of implementation for the current study, as it is reflective of the behavior of the intervention implementer (as opposed to intervention recipient), and so serves as a key source of variability from the program as designed (Berkel, Mauricio, Schoenfelder, & Sandler, 2011). While dosage is also determined by the implementer, the study was not adequately powered for the inclusion of an additional predictor variable at the classroom-level; thus, model over-fitting would have been a significant risk (Myung, 2000). Therefore, another publication by the authors examines the associations between variability in GBG dosage and student outcomes (Ashworth, Panayiotou, Humphrey, & Hennessey, in press).

Cumulative risk. Previous research by the authors (Ashworth & Humphrey, 2018; Ashworth, Humphrey, & Hennessey, under review) identified the student- and school-level risk factors that were significant predictors of baseline behavioral and academic outcomes in the study sample (Table 2). As CR theory states that the number of risk factors is more important than their nature, they were dichotomized (coded as either ‘0’ for absent or ‘1’ for present) and summed for each of the outcomes, creating four CR scores for each child that represented the number of risk factors to which they were exposed. Line graphs were plotted

and calculations of effect size between risk levels were conducted; risk categorizations were then determined based on notable differences in mean scores between risk levels. For instance, while there were similar, small differences in reading point scores between the majority of the risk levels (d=.32-.37), larger differences were evident between risk levels 0 and 1 (d=.43), and 2 and 3 (d=.56). This was consistent with the cumulative risk graph, where an elbow point was visible after exposure to two risk factors. Students were therefore categorized into one of four groups for reading: *no risk, low-risk, medium-risk* and *high-risk.* Similar methods meant that students were categorized into one of three groups for each of the three measures of behavior*: low-risk, medium-risk* and *high-risk* (see Table 3).

Interviews. Data were collected using bespoke semi-structured interview schedules. The schedule acted as a guide to ensure specific topics were addressed, whilst also allowing for unanticipated responses (Galletta, 2013). Questions explored *how* the GBG was implemented (e.g., fidelity, quality, adaptations), *why* it was implemented in this way (i.e., factors affecting implementation), and perceived impact of the GBG, including differential effects. Prompts and probes were utilized where necessary to encourage participants to elaborate on their answers and to clarify unclear responses.

Two interviews were conducted in each year of the trial, at the end of the first and second terms. The first focused more heavily on early implementation, while the second explored perceptions of impact in greater depth. Interviews were conducted by members of the research team with teachers implementing the GBG, and with a member of the SLT in each school, in a private room.

Quantitative Analysis

Multi-level modeling (in MLwiN 2.36) was used to account for the clustered and hierarchical nature of the data (students nested within classes; Twisk, 2006). Prior to analysis, reading and behavior scores were standardized by converting them to z scores, in order to

facilitate interpretation within and across models; this also means that the coefficients reported can be interpreted as effect sizes akin to Cohen's d. Fidelity/quality scores (one for each year of implementation) were converted into binary high/low variables. While no standard criterion for the cut-off for high fidelity has been established, guidelines of 80-100% have been recommended based on previous literature (Perepletchikova & Kazdin, 2005). Thus, teachers scoring above 80% were categorized as implementing with high fidelity/quality, and those below 80% categorized as implementing with low fidelity/quality.

First, two-level models were fitted for each year group, for each of the four outcome variables of interest (disruptive behavior, concentration problem, pro-social behavior, reading attainment). First and second year fidelity/quality was fitted at the class-level and the relevant baseline behavioral/reading scores at the student-level as predictor variables, to establish any "main effect" associations between implementation fidelity/quality and the outcome variables at post-test (following two years of GBG exposure). Second, the relevant risk group categorization was added to the model at the student-level (with the lowest risk group as the reference category), and cross-level interaction terms between fidelity/quality and risk status were specified.

Power. Guidance on power and sample size for multi-level modelling suggests that the main issue is attaining an appropriate sample size at the second level (classroom-level), as the primary aim of the analysis is to test the effects of variables at this level (Snijders, 2005). One such way of establishing this is to calculate the ratio of subjects per variable (SPV). In the present study, the SPV ratio was 60 (60 classes, one variable) which is above the acceptable threshold (Austin & Steyerberg, 2015). Therefore, the classroom-level sample was considered sufficiently large to permit accurate estimation of the coefficients (Austin & Steyerberg, 2015).

Beyond the above, we needed to confirm that there were a satisfactory number of units at the second level of the model (class/teacher) to accommodate the inclusion of our explanatory variable (fidelity/quality) given the expected amount of variance this would likely explain. A recent study of the PATHS curriculum utilized a comparable two-level model with a similar design and sample size (e.g., implementation variables fitted at the teacher/class level, and student outcomes fitted at the child level; Humphrey et al., 2018). The f^2 statistic, a measure of effect size suitable for regression models (Cohen, 1992), was found to be .087 for this model. This suggests that the implementation factors at the second level explained a significant proportion of the class-level variance; therefore, it could be reasonably concluded that 60 units at the second level of our model was more than sufficient to accommodate the inclusion of a single explanatory variable.

Attrition. Twenty-three percent of classes had ceased implementation by the end of the trial. Although they still complied with post-test data collection protocols, this meant that some implementation data were missing. Thus out of 60 possible classrooms, fidelity/quality data were missing for six classes (10%) in the first year and 15 classes (23%) in the second year. There was also attrition over the course of the trial regarding pupil-level outcome data. Three-hundred-and-ninety students (12.6%) left the school during the course of the trial. Teachers failed to provide post-test behavior scores for a further 182 students (5.9%), meaning that behavior scores were missing for 572 students (18.5%). One-hundred-and-seventy-five students (5.7%) were absent on the day of testing for the HGRT, meaning that post-test reading scores were missing for 565 pupils (18.3%).

Missingness was examined through binary logistic regression to identify the variables that predicted partially observed data, and data were found to be missing at random. Hence, multiple imputation (MI) procedures were conducted to maintain the sample size, to reduce the bias associated with attrition, and to allow for the use of techniques designed for complete

data. MI has been found to be suitable for use with samples with up to 60% attrition (Pampaka, Hutcheson, & Williams, 2016). This was conducted in REALCOM-Impute with demographic variables added as auxiliary (where data were fully observed) and response variables. REALCOM-Impute default settings of 1000 iterations, a burn-in of 100, and a refresh of 10 were utilized, in accordance with guidance produced by Carpenter, Goldstein and Kenward (2011) for multi-level imputation with mixed response types.

Qualitative Analysis

A hybrid thematic analysis was undertaken in accordance with Braun and Clarke's (2006) six-phase guide to establish teachers' perceptions of the differential effects of the GBG for at-risk students, and the mechanisms underpinning any effects. NVivo was utilized to manage the process (https://www.qsrinternational.com/nvivo/). Deductive themes were organized by variables identified as risk factors for either disruptive behavior or reading attainment in preceding quantitative analyses (Ashworth, & Humphrey, 2018; Ashworth et al., under review). Thus, there were nine a priori organizing themes included in the analysis regarding students' risk status (school-level behavior, school-level EAL, gender (male), white EAL, looked-after child, familial poverty, neighborhood poverty, younger relative age - summer born). The analysis of the perceived ways in which the GBG affected at-risk students' outcomes was conducted inductively, to allow for unexpected and emergent themes (Nowell et al., 2017).

Results

Quantitative Results

Descriptive statistics are provided in Table 3. Tables 4 and 5 present main effects and Tables 6 and 7 present subgroup effects, for the first and second years of implementation respectively.

First year of implementation. Regarding overall associations between fidelity/quality and teachers' perceptions of students' behavioral outcomes, higher fidelity/quality was not statistically significantly associated with disruptive behavior (β0ij = .024, p = .427), pro-social behavior (β0ij =.157, p = .174), or concentration problems (β0ij = -.158, p = .142) at post-test. However, it was found to be statistically significantly associated with improved reading point scores at the end of the trial. The associated effect size was small (β0ij = .225, p = .011).

Levels of fidelity/quality did not interact with student-level CR status in relation to teachers' perceptions of concentration problems (medium-risk: β0ij = .072; p = .302; high-risk: β0ij = .166, p = .193), pro-social behavior (medium-risk: β0ij = .046, p = .357; high-risk: β0ij = .160, p = .310) or reading attainment (low-risk: β0ij = -.154; p = .106; medium-risk: β0ij = -.201; p = .109; high-risk: β0ij = -.141, p = .244) at post-test. However, high fidelity/quality was found to be statistically significantly associated with higher levels of teacher perceived disruptive behavior at post-test among high-risk pupils; the associated effect size was medium (β0ij = .560, p = .037).

Second year of implementation. Regarding overall associations between fidelity/quality and students' behavioral outcomes, higher fidelity/quality was not statistically significantly associated with disruptive behavior (β0ij = -.176, p = .157), pro-social behavior (β0ij = .179, p = .197), or concentration problems (β0ij = -.213. p = .120) at post-test. However, as in the first year analysis, it was found to be statistically significantly associated with improved reading point scores at the end of the trial. The associated effect size was small (β0ij = .203, p = .042).

Levels of fidelity/quality did not interact with student-level CR status in relation to disruptive behavior (medium-risk: β0ij = .006, p = .484; high-risk: β0ij = -.320, p = .213), concentration problems (medium-risk: β0ij = .143, p = .216; high-risk: β0ij = .367, p = .063),

pro-social behavior (medium-risk: β0ij = .061, p = .342; high-risk: β0ij =.322, p = .209) or reading attainment (low-risk: β0ij = -.165, p = .157; medium-risk: β0ij = -.168, p = .192; high-risk: β0ij = -.180, p = .246) at post-test.

Qualitative Results

Of the nine a priori themes regarding students' risk factors, four were identified in the dataset, namely, gender (male), SEND status, EAL status (although not ethnicity), and both familial and neighborhood deprivation (although these were discussed indiscriminately by teachers and so form one joint organizing theme). In addition, students with behavioral problems and those with low academic attainment were identified inductively. Thus, six subgroups of students who were perceived to experience differential gains from the intervention were identified.

Inductive analysis focusing on the ways in which the GBG affected at-risk students revealed ten emergent outcomes that were perceived to be impacted (e.g., improved social skills). While these were not always directly related to the outcomes of interest in the quantitative analysis (e.g., disruptive behavior, reading attainment), they could be seen to be proximal effects that may indirectly influence those more distal outcomes (e.g., increased engagement may lead to improved attainment in the longer-term). Five GBG elements were also identified as mechanisms perceived to be underpinning these proposed differential effects (e.g., team leadership).

A thematic map was developed (Figure 1) to summarize the relationships between the risk factors and the emergent outcomes, and the GBG elements perceived to be underlying these associations. While there is not scope in the present study to provide an exhaustive description of all mechanisms and outcomes relating to each risk factor (the reader is referred to Ashworth, 2018 for this), examples are provided here of the most prominent findings that best help to explain the quantitative results.

Benefits for at-risk students. Examples are provided below for two groups of students who were perceived to be benefiting from the GBG.

Low ability students and teamwork. Key aspects of the GBG that were perceived to be beneficial to low ability students were the team membership and leadership elements of the game. Some teachers felt that the team leader aspect helped these lower ability students as it increased their "leadership skills" (teacher 6) and gave them "more responsibility and a bit… of a confidence boost" (teacher 9). Others felt that it allowed them to succeed in other areas, increasing their self-esteem and, consequently, their attainment:

> A few who… have the role to be a team leader and actually thriving on that… it gives them ability to show different aspects of themselves…. So while… they might not be in the top group for Maths… they can be in the top team that's winning the Good Behavior Game, and that in themselves gives them self-esteem which then enhances their learning overall. (SLT_2)

However, other teachers commented instead on the team membership element, noting that it improved independence for these students: "I love the team work aspect… because… I've got the lower ability children and… they find it really hard to be an independent learner" (teacher 12).

EAL students and the explicit nature of the GBG. Teachers from two schools discussed the differential effects on students who were classified as EAL. One teacher felt the explicit nature of the GBG during the pre-game stage was beneficial to both their behavior and independent learning:

> With the Good Behaviour Game you go through the task in so much detail, I think they're a lot clearer about what's expected of them, whereas in an ordinary lesson I wouldn't go into as much specifics about what exactly they need to do and I wouldn't check six times that everybody knows what they're doing, so I think that's why their

> behaviour has improved and they're able to work more independently, because… you make it so clear when you're doing the Good Behaviour Game. (teacher 9)

This teacher was still reporting these differential benefits after one year of implementation, commenting on improvements in understanding for these students: "it definitely helps sort of the lower and the EAL children who've struggled to understand, so it's especially good for them I think". They went on to explain how the explicit nature of the game encouraged EAL students to take responsibility for their own work:

> Just because… I can't intervene during the game I need to make absolutely sure that they know what they're doing before they start and they need to take ownership of that as well because it's their responsibility, if they don't understand what they're doing they'll get strikes for the team so they've kind of switched on a little bit more and listening more. (teacher 9)

Difficulties for at-risk students. Examples are provided below for two groups of students who were perceived to be finding participating in the GBG difficult.

Boys and the rigid structure of the GBG. While one teacher commented specifically on the benefits for the boys in their class regarding behavior ("it works really well especially with my boys who can be often quite destructive… and hard work"; teacher 5), most experienced issues with engagement with some of the boys in their class: "trying to engage those two boys in particular was… really tricky" (teacher 1). Teachers noted that sometimes boys' behavior was worse during the GBG as they would "push it", testing the boundaries and structure of the game, purposefully building up infractions to "see what would happen" (teacher 1). One described this in more detail, explaining how one boy in their class "got it right up to four [infractions] and then just corrected his behavior… he was just deliberately looking at the number going up and then when it got to four he stopped, it was quite annoying" (teacher 9).

SEND students and teacher-student interaction. Another aspect of the GBG that several teachers commented on was the lack of teacher-student interaction permitted during the game. Some teachers found that fidelity to this element was a particular issue with their students with an SEND; one explained that not being able to interact with the teacher could be stressful for these students, impacting their wellbeing:

> I don't like having rigidities that you can't give a little bit of support… they do need that little bit of interaction from adults because sometimes they're not able themselves to express how they're feeling… for some children that five minutes… of the game can be really quite stressful… they know nobody **really** is going to come over to them. (teacher 3)

Another felt that this made managing behavior more difficult: "because we're not able to… go to that child and interact with that child, we therefore can't deescalate the situation… the game just doesn't allow us to do so" (teacher 8). As a result of this, some teachers made adaptations to the game in order to overcome these issues. For example, one teacher allowed these students to have one-to-one support from a teaching assistant during the game: "I have one child who has autism… At the beginning, I made [teaching assistant] part of the group" (teacher 14), while another would intervene to provide additional directions when necessary:

> The only change that I've made a couple of times was when I've had to intervene… my children are… SEN so if they ever got stuck [at] any point or… didn't understand what to do then I'd just… intervene in that way… just to give them a little bit more direction or if they're… way off with something that they've… started work on, I'll just kind of point them in the right direction. (teacher 13)

Discussion

The results of this study demonstrate that higher levels of fidelity/quality of GBG, assessed in both the first and second years of implementation, were associated with improved

reading scores for all students. However, fidelity/quality of GBG implementation was not associated with differential gains for reading for students at varying levels of CR exposure. No overall effects of fidelity/quality were found for teacher perceptions of disruptive behavior, concentration problems or pro-social behavior in either year of the trial. However, high fidelity/quality in the first year of implementation was associated with *worsening* disruptive behavior scores for high-risk students. No differential effects were identified for either concentration problems or pro-social behavior.

Six themes were identified in the qualitative dataset; namely, gender (male), SEND status, EAL status, deprivation, students with behavioral problems, and students with low academic attainment. Teachers identified five key GBG elements (team leadership, team membership, the explicit nature of the GBG, lack of teacher-pupil interaction permitted, and the rigid structure of the game) that were theorized to be the mechanisms through which these students' outcomes were influenced. Ten prominent outcomes emerged and related to a broad range of perceived benefits including self-esteem, social skills, and engagement with learning.

Higher Fidelity/Quality, Improved Reading

While no overall effects of GBG delivery were found for reading attainment in the GBG trial (Ashworth et al., in press.; Humphrey et al., 2018), results from the present study suggest that the intervention may improve students' reading attainment when it is implemented with high levels of fidelity/quality. Thus, it appears that simply implementing the GBG is not adequate when aiming to improve students' reading; in order to be successful in achieving these outcomes, it needs to be delivered well and with close adherence to the program manual. While these findings are in line with the only other study of the GBG to examine the moderating influence of implementation variability (Ialongo et al., 1999), the applicability of Ialongo et al.'s findings could not be assumed. It was important that more current research was conducted in this area, specifically in a UK context. However, it is clear

that the present study supports longstanding concerns in the literature regarding the importance of examining implementation when testing the efficacy of interventions (Durlak, 2016).

It is noteworthy that the program's logic model indicates that attainment is a secondary outcome that is improved via increased attention and on-task behavior, and a decrease in disruptive behaviors (Chan et al., 2012; Ford et al., 2014). However, while high levels of fidelity/quality were associated with improved reading scores for students, no effects were found in the present study for any of the three aspects of behavior (Ashworth et al., under review), regardless of the way the GBG was implemented. This suggests that the improvements in students' reading were mediated by something not measured in the present study. In light of these findings, the GBG logic model may benefit from further development, in order to more accurately represent the mechanisms through which it influences academic achievement, particularly with regards to reading attainment.

Some possible explanations for the role of fidelity and quality in improving students' reading attainment are evident in our qualitative findings. For instance, some teachers suggested that the explicit nature of the GBG (e.g., clear rules, detailed instructions prior to the game beginning) helped at-risk students to understand the task and take responsibility for their work; this in turn better equipped them to complete the tasks given to them, and thus enhanced their learning and attainment. Other teachers reported improvements for low ability students in terms of self-esteem and confidence due to both the teamwork elements of the game and the increased support from peers (and reduced reliance on the teacher necessitated by the removal of teacher-student interaction during gameplay). Therefore, it is likely that in order for gains in academic achievement to be evidenced, teachers needed to closely adhere to these elements of the GBG. However, it is important to note that comments regarding the impact of the game on low ability students were mixed, with some teachers reporting

difficulties for certain groups. Thus, it is possible that there may have been other factors affecting outcomes for these students.

Furthermore, all of the more proximal outcomes identified in our qualitative analysis could be associated with improved academic outcomes in the longer-term. For example, self-esteem has previously been found to be directly related to academic achievement (Marsh & Martin, 2011). Indeed, in a concurrent paper by the authors (Ashworth et al., in press) reading attainment was found to improve only at one-year post-intervention follow-up, and only when the GBG was implemented with at least moderate levels of dosage, suggesting that the impact on attainment is not always immediate. Thus, it may be worth exploring these proximal outcomes more explicitly when re-examining the program's logic model, as they could potentially be underpinning improvements in academic achievement, as opposed to the behavioral outcomes currently cited as mechanisms.

Higher Fidelity/Quality, Worsening Disruptive Behavior for High-Risk Students

Contrary to the finding that high fidelity/quality of GBG delivery was associated with overall improvements in reading attainment, no such association was found for any of the three aspects of teacher-perceived behavior measured in the present study. When considered alongside the null main effects identified in the wider trial regarding behavior (Ashworth et al., under review; Humphrey et al., 2018), it appears that these results were not due to poor delivery of the GBG. Instead, these findings suggest that the GBG was simply not effective at improving behavioral outcomes for students in English schools, regardless of the way it was implemented.

However, not only did the GBG have no significant benefit for teachers' perceptions of students' behavioral outcomes, high-risk students' disruptive behavior scores *worsened* when the GBG was delivered with high fidelity/quality. Whilst no GBG research has previously examined the association between implementation variability and outcomes for

students at varying levels of risk, this is contrary to the outcomes that have been hypothesized. For instance, Muthén et al. (2002) posited that the GBG would not be intensive enough to have an impact on the high-risk group. Furthermore, it is typically assumed that closer adherence to the prescribed components of an intervention will result in greater improvements in outcomes (Durlak, 2016), and previous GBG research generally evidences greater gains for at-risk students, regardless of implementation variability (e.g., Kellam et al., 2011).

Similarly to the outcomes regarding reading attainment, the qualitative data helps to explain this seemingly incongruous finding. Teachers reported that several elements of the GBG such as the rigid structure, the removal of teacher-student interaction, and lack of available support was problematic for certain groups of at-risk students. This in turn meant that these students became disengaged during the game and that teachers could not intervene to de-escalate situations, resulting in an increase in behavioral issues. In an attempt to overcome their concerns regarding the perceived negative effects of the game on at-risk students' outcomes, teachers made a variety of adaptations: they provided these students with additional feedback, interacted with them during the game, and allowed them to have a one-to-one support. In fact, teachers only typically reported benefits of the GBG for at-risk students *if* they made adaptations. This suggests that certain core elements of the GBG outlined in the manual as critical to successful implementation were actually considered to be detrimental to at-risk students. As the teachers in the high fidelity/quality group were likely those who adhered to the manual and made fewer of these seemingly necessary adaptations, this provides an explanation as to why high fidelity/quality appeared to have negative associations with high-risk students' behavior.

Alternatively, it is possible that other factors not measured in the present study, such as participant responsiveness or reach, may have been influencing the results. For instance, in

an evaluation of the PATHS curriculum, participant reach was the only factor of implementation significantly associated with outcome variability across all of the analyses (Humphrey et al., 2015). It may be that high-risk students who were in classrooms where the GBG was implemented with strict adherence to the manual may have found that there was not enough support available to meet their needs and so responded poorly to the intervention. Conversely, it is possible that teachers utilized the time that the GBG was played to withdraw high-risk students for other more targeted interventions or nurture groups (Askell-Williams, 2015). Thus, while these pupils were technically 'in' the high fidelity/quality classrooms, they were not necessarily present for the intervention. Indeed, while reach was high (95.6%), this does suggest that, a small proportion (4.4%) were not present for GBG delivery. Although it cannot be determined that those that were missing were the high-risk students, it is certainly a possibility. Thus, future research should seek to examine both aspects of implementation related to the deliverer and the participants.

Implications

As first proposed by Durlak and DuPre (2008), implementation does indeed matter, and the findings from the present study highlight the importance of monitoring it when evaluating the effectiveness of interventions, and its moderating role in intervention outcome variability. Furthermore, while the initial quantitative results regarding behavior appeared somewhat incongruous with the GBG logic model and previous literature in the field, the qualitative findings provided useful possible explanations for these. Thus, future evaluations of school-based interventions should seek to incorporate mixed-methods IPEs as standard practice when examining the effectiveness of these programs in order to guard against Type III errors.

It is also apparent that while high fidelity/quality is important with regards to reading outcomes, strict adherence to the manual may not necessarily always be the answer. These

results have important implications for schools seeking effective universal interventions, as they suggest that a "one size fits all" approach regarding implementation may actually be hindering progress for certain groups of students (NEA, 2014). Thus, schools will need to be cautious regarding the interventions they choose to implement, and the ways in which they do this. The inflexibility of prescriptive interventions may fail to account for the varying needs between groups of students.

Although it appears that high fidelity/quality needs to be maintained in order to benefit reading outcomes, further research is needed to identify the adaptations that teachers could make for specific subgroups of students, in order to ensure they are also perceived to be benefiting from the intervention regarding their behavioral outcomes. The 'fidelity-adaptation' debate is ongoing (Lendrum & Humphrey, 2012), with tensions between intervention developers' desire for strict adherence to the program, and implementers' wishes to adapt the intervention to suit the context. It has been suggested that interventions that are not flexible enough to meet the needs of the context are at risk of failing (Greenberg, Domitrovich, Graczyk, & Zins, 2005), and it appears that this may be the case in regard to high-risk students' disruptive behavior. Therefore, intervention developers may want to work on establishing the adaptations in line with the program logic model that can be implemented, to ensure that these interventions are truly universal. Lendrum and Humphrey (2012) distinguish between adaptations that are considered to be modifications to existing components, and those that are additions to an intervention, arguing that these do not equate to a lack of fidelity. Therefore, future research could explore the additions that can be made to the GBG, as opposed to modifications that interfere with the intervention's underlying theory of change; this may help to ensure that fidelity to the critical components remains high in order to benefit reading attainment, whilst also guarding against any perceived detrimental behavioral effects for at-risk students.

The findings from the present study highlight not only the importance of examining the association between implementation variability and differential subgroup gains for students, but also provide support for the utility of CR exposure as a viable means through which to examine these differential effects. As Greenberg and Abenavoli (2017) noted, the preoccupation with main effects in research has led to an underappreciation of the natural heterogeneity that exists in populations receiving universal interventions, meaning important effects are missed. Had only main effects analyses been conducted, the potentially detrimental impact of the intervention on high-risk students' disruptive behavior would not have been identified. However, traditional subgroup analyses (e.g., those based on a single risk factor) would also have failed to identify this. Indeed, examining subgroup effects of preventive interventions based on one risk factor in isolation does not provide an accurate representation of the individual differences and multitude of factors influencing a child's development. This is in line with ecological systems theory underpinning CR research, which suggests that children do not develop in a vacuum, but are shaped by many different factors in a variety of domains (Bronfenbrenner, 1986). Therefore, CR indices are arguably a superior tool to use when examining differential gains in the context of universal interventions. Thus, future research should seek to ensure differential effects for multiple subgroups are examined when testing the efficacy of an intervention, in order to ensure results are not biased against the kinds of effects universal interventions may yield.

Finally, it is important to note that the present study was highly exploratory, particularly regarding the quantitative subgroup analysis, and future research should seek to replicate and extend these findings. Furthermore, the design utilized means that the sample size was reduced, as only schools in the GBG arm of the trial could be included in analyses for the purposes of the present study. However, other forms of analysis that incorporate participants in both trial arms, such as complier average causal effect estimation (CACE),

should be utilized in future research, in order to provide a more robust estimation of the effects of intervention compliance on students' outcomes. This is particularly pertinent considering the proportion of teachers who ceased implementation over the course of the trial.

Limitations

Although observations are considered to be the most valid method for assessing implementation (Humphrey et al., 2016) and there were no concerns regarding inter-rater reliability or researcher effects, only one observation was conducted for each teacher. This therefore only provides a single snapshot of implementation, and does not account for contextual factors that may have been influencing implementation on the day of the observation. Repeated observations over multiple time points would have been desirable to provide a more representative average rating. Additionally, social desirability may have influenced the way that teachers delivered the GBG when they were being observed.

It is recommended that all eight aspects of implementation should be included in analyses, in order to gain a more complete picture (Durlak, 2015; Durlak & DuPre, 2008; Humphrey et al., 2016). However, as only fidelity/quality were assessed in this study, some important factors influencing outcomes may have been overlooked.

Furthermore, although fidelity/quality was originally a continuous variable, it was dichotomized prior to analysis, utilizing an external cut-off of 80% to represent high fidelity/quality. While this external cut-off is often recommended (Savignac & Dunbar, 2014), this does cause some information to be lost, and leads to uneven sample sizes. Indeed, sample sizes in the present study were very small for some at-risk subgroups due both to the fidelity/quality cut-off and the method utilized to determine the risk group categorizations. Thus it is possible that results in the subgroup analyses may have been skewed or spurious, and so should be interpreted with caution. Indeed, the exploratory nature of this study should be acknowledged and results found here require replication.

Finally, it is noteworthy that the primary measure of student behavior was a checklist completed by the students' teachers. While it was not possible to observe behavior independently due to the scale of the study, there are potential concerns with using an indirect approach to such measurement. In particular, it is possible that this approach contributed to the finding that higher levels of fidelity were associated with higher levels of perceived disruptive behavior at post-test, as teachers who had higher levels of fidelity may have expected greater changes in those disruptive students' behaviors and thus had a response bias when completing the checklist post-test. Indeed, previous studies have shown low correlations between different raters and observations, and an examination of the TOCA-C indicated that teachers' perceptions of their environment influenced the scores they provided for their students (Pas & Bradshaw, 2014). Furthermore, while we have hypothesized that improvements in academic outcomes were not mediated by attention or on-task behavior, as these behaviors were measured by a checklist, it is possible that those behaviors did increase, but the measurement system used did not capture this. Future research may wish to utilize a mixed-methods study to include some direct observation procedures combined with the use of checklists to more directly index student behavior and validate what the checklists are capturing.

Conclusions

The present study is, to the authors' knowledge, the first to examine the interaction between levels of implementation and participant risk status as predictors of intervention outcomes. Furthermore, the study has advanced the examination of differential gains, an area highlighted as a priority for future research in Durlak et al.'s (2011) meta-analysis, by utilizing CR exposure as a subgroup marker. We also contribute to the evidence base in this area by exploring *why* students at different levels of risk exposure may respond differently to varying levels of implementation. We conclude that while higher levels of fidelity/quality are

associated with improved overall reading attainment, strict adherence to the manual is not always beneficial with regards to high-risk students' disruptive behavioral outcomes.

References

Ashworth, E. (2018). *Differential effects of the Good Behaviour Game on pupils' school functioning: Cumulative risk exposure as a moderator of intervention outcomes* (Doctoral dissertation). Retrieved from https://www.research.manchester.ac.uk/portal/en/theses/search.html

Ashworth, E. & Humphrey, N. (2018). More than the sum of its parts: Cumulative risk effects on school functioning in middle childhood. *British Journal of Educational Psychology*. https://doi.org/10.1111/bjep.12260

Ashworth, E., Demkowicz, O., Lendrum, A., & Frearson, K. (2018). Coaching Models of School-Based Prevention and Promotion Programmes: A Qualitative Exploration of UK Teachers' Perceptions. *School Mental Health*. https://doi.org/10.1007/s12310-018-9282-3

Ashworth, E., Humphrey, N., & Hennessey, A. (under review). Game over? No main or subgroup effects of the Good Behavior Game in a randomized trial in English primary schools. *Journal of Research on Educational Effectiveness*.

Ashworth, E., Panayiotou, M., Humphrey, N., & Hennessey, A. (in press). Game on - complier average causal effect estimation reveals sleeper effects on academic attainment in a randomized trial of the Good Behavior Game. *Prevention Science*.

Askell-Williams, H. (2015). *Transforming the future of learning with educational research*. Hersey, PA: Information Science Reference.

Askell-Williams, H., Dix, K., Lawson, M., & Slee, P. (2012). Quality of implementation of a school mental health initiative and changes over time in students' social and emotional competencies. *School Effectiveness and School Improvement, 24*, 357–381. https://doi.org/10.1080/09243453.2012.692697

Austin, P., & Steyerberg, E. (2015). The number of subjects per variable required in linear

regression analyses. *Journal of Clinical Epidemiology*, *68*, 627–636. https://doi.org/10.1016/j.jclinepi.2014.12.014

Barrish, H., Saunders, M., & Wolf, M. (1969). Good Behavior Game: Effects of individual contingencies for group consequences on disruptive behavior in a classroom. *Journal of Applied Behavior Analysis*, *2*, 119–124. https://doi.org/10.1901/jaba.1969.2-119

Berkel, C., Mauricio, A., Schoenfelder, E., & Sandler, I. (2011). Putting the pieces together: An integrated model of program implementation. *Prevention Science*, *12*, 23–33. https://doi.org/10.1007/s11121-010-0186-1

Bonell, C., Fletcher, A., Morton, M., Lorenc, T., & Moore, L. (2012). Realist randomised controlled trials: A new approach to evaluating complex public health interventions. *Social Science and Medicine*, *75*, 2299–2306. https://doi.org/10.1016/j.socscimed.2012.08.032

Bradshaw, C., Waasdorp, T., & Leaf, P. J. (2015). Examining variation in the impact of school-wide positive behavioral interventions and supports: Findings from a randomized controlled effectiveness trial. *Journal of Educational Psychology*, *107*, 546–557. https://doi.org/10.1037/a0037630

Bronfenbrenner, U. (1986). Ecology of the family as a context for human development: Research perspectives. *Developmental Psychology*, *22*, 723–742. https://doi.org/10.1037/0012-1649.22.6.723

Bruhn, A., Hirsch, S., & Lloyd, J. (2015). Treatment integrity in school-wide programs: A review of the literature (1993–2012). *Journal of Primary Prevention*, *36*, 335–349. https://doi.org/10.1007/s10935-015-0400-9

Carpenter, J., Goldstein, H., & Kenward, M. (2011). REALCOM-IMPUTE software for multilevel multiple imputation with mixed response types. *Journal of Statistical Software*, *45*, 1–14. https://doi.org/http://dx.doi.org/10.18637/jss.v045.i05

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

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Devine, A., Soltész, F., Nobes, A., Goswami, U., & Szucs, D. (2013). Gender differences in developmental dyscalculia depend on diagnostic criteria. *Learning and Instruction*, *27*, 31–39. https://doi.org/10.1016/j.learninstruc.2013.02.004

DfE. (2015). *Schools, pupils, and their charateristics: January 2015*. London, UK: DfE.

Dolan, L., Kellam, S., Hendricks Brown, C., Werthamer-Larsson, L., Rebok, G., Mayer, L., … Wheeler, L. (1993). The short-term impact of two classroom-based preventive interventions of aggressive and shy behaviors and poor achievement. *Journal of Applied Developmental Psychology*, *14*, 317–345.

Domitrovich, C., Pas, E., Bradshaw, C., Becker, K., Keperling, J., Embry, D., & Ialongo, N. (2015). Individual and School Organizational Factors that Influence Implementation of the PAX Good Behavior Game Intervention. *Prevention Science*, *16*, 1064–1074. https://doi.org/10.1007/s11121-015-0557-8

Donaldson, J., Vollmer, T., Krous, T., Downs, S., & Berard, K. (2011). An evaluation of the Good Behavior Game in kindergarten classrooms. *Journal of Applied Behavior Analysis*, *44*, 605–609. https://doi.org/10.1901/jaba.2011.44-605

Durlak, J. (2015). Studying program implementation is not easy but it is essential. *Prevention Science*, *16*, 1123–1127. https://doi.org/10.1007/s11121-015-0606-3

Durlak, J. (2016). Programme implementation in social and emotional learning: Basic issues and research findings. *Cambridge Journal of Education*, *46*, 333–345. https://doi.org/10.1080/0305764X.2016.1142504

Durlak, J., & DuPre, E. (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*, 327–350. https://doi.org/10.1007/s10464-008-9165-0

Durlak, J., Weissberg, R., Dymnicki, A., Taylor, R., & Schellinger, K. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child Development*, *82*, 405–432. https://doi.org/10.1111/j.1467-8624.2010.01564.x

EEF. (2013). *Pre-testing in EEF evaluations*.

Farrell, A., Henry, D., & Bettencourt, A. (2013). Methodological challenges examining subgroup differences: Examples from universal school-based youth violence prevention trials. *Prevention Science*, *14*, 121–133. https://doi.org/http://doi.org/10.1007/s11121-011-0200-2

Ford, C., Keegan, N., Poduska, J., Kellam, S., & Littman, J. (2014). *Implementation manual*. Washington, DC: American Institutes for Research.

Galletta, A. (2013). *Mastering the semi-structured interview and beyond: From research design to analysis and publication*. New York: New York University Press.

Gerard, J., & Buehler, C. (1999). Multiple risk factors in the family environment and youth problem behaviors. *Journal of Marriage and the Family*, *61*, 343–361.

Greenberg, M., & Abenavoli, R. (2017). Universal Interventions: Fully Exploring Their Impacts and Potential to Produce Population-Level Impacts. *Journal of Research on Educational Effectiveness*, *10*, 40–67. https://doi.org/10.1080/19345747.2016.1246632

Greenberg, M., Domitrovich, C., Graczyk, P., & Zins, J. (2005). The study of implementation in school-based preventive interventions: Theory, research, and practice. *Promotion of Mental Health and Prevention of Mental and Behavior Disorders 2005 Series V3*.

Hagermoser Sanetti, L., Dobey, L., & Gallucci, J. (2014). Treatment integrity of interventions with children in School Psychology International from 1995-2010. *School Psychology International*, *35*, 370–383. https://doi.org/10.1177/0143034313476399

Hagermoser Sanetti, L., & Fallon, L. (2011). Treatment integrity assessment: How estimates of adherence, quality, and exposure influence interpretation of implementation. *Journal of Educational and Psychological Consultation*, *21*, 209–232. https://doi.org/10.1080/10474412.2011.595163

Hallgren, K. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*, 23–34. https://doi.org/10.1016/j.biotechadv.2011.08.021.Secreted

Harachi, T., Abbott, R., Catalano, E., Haggerty, K., & Fleming, C. (1999). Opening the black box: Using Process evaluation measures to assess implementation and theory building. *American Journal of Community Psychology*, *27*, 711–731.

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. https://doi.org/10.1080/1754730X.2009.9715704

Humphrey, N., Barlow, A., & Lendrum, A. (2017). Quality matters: Implementation moderates student outcomes in the PATHS curriculum. *Prevention Science*, *19*, 197–208. https://doi.org/10.1007/s11121-017-0802-4

Humphrey, N., Barlow, A., Wigelsworth, M., Lendrum, A., Pert, K., Joyce, C., … Turner, A. (2015). *Promoting Alternative Thinking Strategies (PATHS): Evaluation report and executive summary*. London, UK: EEF.

Humphrey, N., Hennessey, A., Ashworth, E., Frearson, K., Petersen, K., Wo, L., … Pampaka, M. (2018). *Good Behaviour Game: Evaluation report and executive summary*. London,

Uk.

Humphrey, N., Lendrum, A., Ashworth, E., Frearson, K., Buck, R., & Kerr, K. (2016). *Implementation and process evaluation (IPE) for interventions in educational settings: A synthesis of the literature*. *EEF*. London, UK. https://doi.org/10.1017/CBO9781107415324.004

Ialongo, N., Werthamer, L., Kellam, S., Brown, C., Wang, S., & Lin, Y. (1999). Proximal impact of two first-grade preventive interventions on the early risk behaviors for later substance abuse, depression, and antisocial behavior. *American Journal of Community Psychology*, *27*, 599–641. https://doi.org/10.1023/A:1022137920532

Jones, S., Brown, J., & Aber, J. (2011). Two-year impacts of a universal school-based social-emotional and literacy intervention: An experiment in translational developmental research. *Child Development*, *82*, 533–554. https://doi.org/10.1111/j.1467-8624.2010.01560.x

Kellam, S., Mackenzie, A., Brown, C., Poduska, J., Wang, W., Petras, H., & Wilcox, H. (2011). The Good Behavior Game and the future of prevention and treatment. *Addiction Science & Clinical Practice*, *6*, 73–84.

Koth, C., Bradshaw, C., & Leaf, P. (2009). Teacher observation of classroom adaptation-checklist: Development and factor structure. *Measurement and Evaluation in Counseling and Development*, *42*, 15–30. https://doi.org/10.1177/0748175609333560

Kourkounasiou, M., & Skordilis, E. (2014). Validity and reliability evidence of the TOCA-C in a sample of Greek students. *Psychological Reports*, *115*, 766–783.

Lastrapes, R. E. (2013). Using the Good Behavior Game in an inclusive classroom. *Intervention in School and Clinic*, *49*, 225–229. https://doi.org/10.1177/1053451213509491

Lendrum, A., & Humphrey, N. (2012). The importance of studying the implementation of

interventions in school settings. *Oxford Review of Education, 38,* 635–652. https://doi.org/10.1080/03054985.2012.734800

Marsh, H., & Martin, A. (2011). Academic self-concept and academic achievement: Relations and causal ordering. *British Journal of Educational Psychology, 81,* 59–77. https://doi.org/10.1348/000709910X503501

Muthén, B., Brown, C. H., Booil, K., Khoo, S., Yang, C., Wang, C., … Liao, J. (2002). General growth mixture modeling for randomized preventive interventions. *Biostatistics, 3,* 459–475.

Myung, I. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, *44,* 190–204. https://doi.org/10.1006/ jmps.1999.1283.

NEA. (2014). *Positive behavioral interventions and supports: A multi-tiered framework that works for every student*.

O'Donnell, C. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K-12 curriculum intervention research. *Review of Educational Research, 78,* 33–84. https://doi.org/10.3102/0034654307313793

Pampaka, M., Hutcheson, G., & Williams, J. (2016). Handling missing data: Analysis of a challenging data set using multiple imputation. *International Journal of Research & Method in Education, 39,* 19–37. https://doi.org/10.1080/1743727X.2014.979146

Pas, E., & Bradshaw, C. (2014). What affects teacher ratings of student behaviors? The potential influence of teachers' perceptions of the school environment and experiences. *Prevention Science*, *15,* 940–950. https://doi.org/10.1007/s11121-013-0432-4

Perepletchikova, F., & Kazdin, A. (2005). Treatment integrity and therapeutic change: Issues and research recommendations. *Clinical Psychology: Science and Practice*, *12,* 365–383. https://doi.org/10.1093/clipsy/bpi045

Rutter, M. (1979). Maternal deprivation, 1972-1978: New findings, new concepts, new

approaches. *Child Development, 50*, 283–305.

Sameroff, A., Gutman, L., & Peck, S. (2003). Adaptation among youth facing multiple risks: Protective research findings. In *S. Luthar (Ed). Resilience and Vulnerability* (pp. 364–391). Cambridge, UK: Cambridge University Press. https://doi.org/10.1176/appi.ajp.162.8.1553-a

Savignac, J., & Dunbar, L. (2014). *Guide on the implementation of evidence-based programs: What do we know so far?*

Snijders, T. (2005). Power and sample size in multilevel modeling. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of Statistics in Behavioral Science* (pp. 1570–1573). Chicester: Wiley. https://doi.org/10.1093/jac/41.5.513

Tabachnick, B., & Fidell, L. (2015). *Using multivariate statistics* (6th ed.). Essex, UK: Pearson Education Limited.

The Multisite Violence Prevention Project. (2008). The multisite violence prevention project: Impact of a universal school-based violence prevention program on social-cognitive outcomes. *Prevention Science, 9*, 231–244. https://doi.org/10.1007/s11121-008-0101-1

Twisk, J. (2006). *Applied multilevel analysis*. Cambridge, UK: Cambridge University Press.

Wigelsworth, M., Lendrum, A., Oldfield, J., Scott, A., ten Bokkel, I., Tate, K., & Emery, C. (2016). The impact of trial stage, developer involvement and international transferability on universal social and emotional learning programme outcomes: A meta-analysis. *Cambridge Journal of Education, 46*, 347–376. https://doi.org/10.1080/0305764X.2016.1195791

Table 1

Demographic data for the school- and student-level samples

Quantitative strand				
		Average school-level sample (N=77)	Average student-level sample (N=3,084)	
Size – number of pupils on roll		298.2	-	
Sex – % of male students		-	50.4	
FSM – % of pupils eligible for FSM		27.6	27.4	
EAL – % of pupils speaking EAL		22	26.2	
Ethnic Minority – % of ethnic minority pupils		32.4	32.8	
SEND – % of pupils with SEND		20.9	23.1	
Qualitative strand				
Teacher code	Trial year	Fidelity/quality %	Years qualified (baseline)	Gender
1	1	75.60	1	M
2	2	67.65	12	F
3	2		8	F
SLT_1	SLT	-	-	F
4	1	60.83	8	F
5	1&2		0	F
6	2	51.86	1	M
SLT_2	SLT	-	-	F
7	1	63.91	2	F

8	2	40.85		F
9	1&2	64.75	0	F
SLT_3	SLT	-	-	F
10	1	68.33		M
SLT_4	SLT	-	-	F
SLT_5	SLT	-	-	F
11	1	77.07	2	F
12	1	77.07	3	F
13	2	74.71	15	M
14	2	67.76	36	F
SLT_6	SLT	-	-	F

Note.

FSM – free school meals

EAL – English as an additional language

SEND – special educational needs and disabilities

SLT – senior leadership team

Table 2

Cumulative risk index development

Risk factor	Disruptive behavior	Concentration problems	Pro-social behavior	Reading attainment
School-level				
High urbanicity		✓		
High urbanicity		✓		
% EAL students	✓		✓	✓
% students conduct problems	✓	✓	✓	
Student-level				
Male gender	✓	✓	✓	✓
Summer-born		✓		✓
FSM eligible	✓	✓	✓	✓
SEND	✓	✓	✓	✓
Looked-after child	✓	✓	✓	
White EAL				✓
High neighborhood deprivation				✓

Note.

FSM – free school meals

EAL – English as an additional language

SEND – special educational needs and disabilities

Table 3

Descriptive data and mean (standard error) fidelity/quality and outcome scores

	N risk factors	N students	Pre-test	Post-test
Disruptive behavior			1.66 (.022)	1.74 (.025)
Low-risk	0+1	762	1.46 (0.21)	1.59 (0.03)
Medium-risk	2+3	644	1.91 (0.37)	1.89 (0.04)
High-risk	4+	82	2.37 (0.10)	2.01 (0.11)
Concentration problems			2.57 (.033)	2.54 (.033)
Low-risk	0+1	285	1.94 (.050)	2.06 (.061)
Medium-risk	2+3	837	2.58 (.037)	2.47 (.043)
High-risk	4+5	303	3.33 (.064)	3.08 (.193)
Pro-social behavior			4.94 (.025)	4.81 (.027)
Low-risk	0+1	782	5.16 (.027)	4.96 (.035)
Medium-risk	2+3	674	4.66 (.036)	4.67 (.042)
High-risk	4+	88	4.25 (.093)	4.53 (.111)
Reading attainment			15.31(0.10)	32.47(0.29)
No risk	0	215	17.51 (.20)	37.17 (.66)
Low-risk	1+2	803	15.80 (.12)	34.20 (.37)
Medium-risk	3	270	13.65 (.26)	28.65 (.70)
High-risk	4+	125	12.10 (.36)	24.57 (1.00)
Fidelity/quality %	N classes year one	N classes year two	Trial year one	Trial year two
Low	43	36	65.88 (.32)	67.03 (.34)
High	11	9	84.95 (.18)	83.58 (.23)

Table 4

Main effects of implementation fidelity/quality on student outcomes: Trial year one

	Disruptive behavior β0ij = -1.407(0.082)			Concentration problems β0ij = -1.439(0.087)			Pro-social behavior β0ij = -2.738(0.173)			Reading attainment β0ij = -3.299(0.092)		
	β	SE	p value	β	SE	p value	β	SE	p value	β	SE	p value
Class- level	0.121	0.029	<.001**	0.162	0.037	<.001**	0.212	0.048	<.001**	0.064	0.016	<.001**
Fidelity/quality (if high)	0.024	0.129	.427	-0.158	0.146	.142	0.157	0.166	.174	0.225	0.095	.011*
Student- level	0.550	0.024	<.001**	0.504	0.022	<.001**	0.608	0.027	<.001**	0.350	0.015	<.001**
Baseline:												
Disruptive	0.876	0.032	<.001**	-	-	-	-	-	-	-	-	-
Concentration	-	-	-	0.575	0.021	<.001**	-	-	-	-	-	-
Pro-social	-	-	-	-	-	-	0.539	0.031	<.001**	-	-	-
Reading	-	-	-	-	-	-	-	-	-	0.210	0.005	<.001**

Table 5

Main effects of implementation fidelity/quality on student outcomes: Trial year two

	Disruptive behavior			Concentration problems			Pro-social behavior			Reading attainment		
	β0ij = -1.469(0.093)			β0ij = -1.519(0.093)			β0ij = -2.773(0.188)			β0ij = -3.249(0.099)		
	β	SE	p value	β	SE	p value	β	SE	p value	β	SE	p value
Class- level	0.153	0.039	<.001**	0.169	0.042	<.001**	0.234	0.056	<.001**	0.068	0.018	<.001**
Fidelity/quality (if high)	-0.176	0.173	.157	-0.213	0.179	.120	0.179	0.208	.197	0.203	0.115	.042*
Student- level	0.537	0.026	<.001**	0.502	0.024	<.001**	0.573	0.028	<.001**	0.352	0.017	<.001**
Baseline:												
Disruptive	0.946	0.036	<.001**	-	-	-	-	-	-	-	-	-
Concentration	-	-	-	0.590	0.022	<.001**	-	-	-	-	-	-
Pro-social	-	-	-	-	-	-	0.559	0.034	<.001**	-	-	-
Reading	-	-	-	-	-	-	-	-	-	0.206	0.006	<.001**

Table 6

Subgroup effects of implementation fidelity/quality on student outcomes: Trial year one

	Disruptive behavior			Concentration problems			Pro-social behavior			Reading attainment		
	β0ij = -1.432(0.088)			β0ij = -1.528(0.101)			β0ij = -2.361(0.189)			β0ij = -3.386(0.125)		
	β	SE	p value	β	SE	p value	β	SE	p value	β	SE	p value
Class- level	0.139	0.033	<.001**	0.167	0.039	<.001**	0.243	0.054	<.001**	0.066	0.017	<.001**
Fidelity/quality (if high)	-0.043	0.149	.387	-0.182	0.183	.162	0.127	0.187	.500	0.340	0.141	.010**
Student- level	0.539	0.024	<.001**	0.480	0.022	<.001**	0.586	0.026	<.001**	0.344	0.015	<.001**
Baseline:												
Disruptive	0.840	0.034	<.001**	-	-	-	-	-	-	-	-	-
Concentration	-	-	-	0.540	0.023	<.001**	-	-	-	-	-	-
Pro-social	-	-	-	-	-	-	0.498	0.032	<.001**	-	-	-
Reading	-	-	-	-	-	-	-	-	-	0.210	0.006	<.001**

Risk group:												
No-risk	-	-	-	-	-	-	-	-	-	◊	◊	◊
Low-risk	◊	◊	◊	◊	◊	◊	◊	◊	◊	0.155	0.063	.009*
Med-risk	0.194	0.065	.002*	0.129	0.073	.042*	-0.339	0.068	<.001**	0.107	0.081	.093
High-risk	0.101	0.136	.230	0.369	0.096	<.001**	-0.455	0.140	.001**	0.015	0.101	.441
High fidelity/quality* low-risk	-	-	-	-	-	-	-	-	-	-0.154	0.122	.106
High fidelity/quality* med-risk	0.090	0.120	.228	0.072	0.138	.302	0.046	0.125	.357	-0.201	0.161	.109
High fidelity/quality* high-risk	0.560	0.308	.037*	0.166	0.190	.193	0.160	0.320	.310	-0.141	0.202	.244

Table 7

Subgroup effects of implementation fidelity/quality on student outcomes: Trial year 2

	Disruptive behavior β0ij = -1.518(0.096)			Concentration problems β0ij = -1.597(0.106)			Pro-social behavior β0ij = -2.376(0.202)			Reading attainment β0ij = -3.346(0.140)		
	β	SE	p value	β	SE	p value	β	SE	p value	β	SE	p value
Class- level	0.176	0.043	<.001**	0.169	0.043	<.001**	0.243	0.058	<.001**	0.070	0.019	<.001**
Fidelity/quality (if high)	-0.175	0.198	.191	-0.330	0.232	.081	0.151	0.224	.252	0.326	0.179	.038*
Student- level	0.526	0.025	<.001**	0.471	0.024	<.001**	0.553	0.026	<.001**	0.350	0.018	<.001**
Baseline:												
Disruptive	0.916	0.037	<.001**	-	-	-	-	-	-	-	-	-
Concentration	-	-	-	0.561	0.025	<.001**	-	-	-	-	-	-
Pro-social	-	-	-	-	-	-	0.510	0.035	<.001**	-	-	-
Reading	-	-	-	-	-	-	-	-	-	0.206	0.007	<.001**

Risk group:												
No-risk	-	-	-	-	-	-	-	-	-	◊	◊	◊
Low-risk	◊	◊	◊	◊	◊	◊	◊	◊	◊	0.168	0.070	.010*
Med-risk	0.200	0.065	.002**	0.086	0.076	.132	-0.308	0.066	<.001**	0.078	0.090	.195
High-risk	0.304	0.161	.033	0.328	0.102	.001*	-0.581	0.160	<.001**	0.073	0.111	.257
High fidelity/quality* low-risk	-	-	-	-	-	-	-	-	-	-0.165	0.162	.157
High fidelity/quality* med-risk	0.006	0.145	.484	0.143	0.180	.216	0.061	0.149	.342	-0.168	0.191	.192
High fidelity/quality* high-risk	-0.320	0.398	.213	0.367	0.235	.063	0.322	0.394	.209	-0.180	0.260	.246