Ashworth, E, Humphrey, N and Hennessey, A

Game Over? No Main or Subgroup Effects of the Good Behavior Game in a Randomized Trial in English Primary Schools

Ashworth, E, Humphrey, N and Hennessey, A (2020) 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. ISSN 1934-5739

Abstract

The current study aimed to examine the impact of a universal, school-based intervention, the Good Behavior Game (GBG), on children's behavior, and to explore any subgroup moderator effects among children at varying levels of cumulative risk (CR) exposure. A two-year cluster-randomized controlled trial was conducted comprising 77 primary schools in England. Teachers in intervention schools delivered the GBG, while their counterparts in control schools continued their usual provision. Behavior (specifically disruptive behavior, concentration problems, and pro-social behavior) was assessed via the checklist version of the Teacher Observation of Classroom Adaptation (TOCA-C). A CR index was calculated by summing the number of risk factors to which each child was exposed. Multi-level models indicated that no main or subgroup effects were evident. These findings were largely insensitive to the modeling of CR, although a small intervention effect on disruptive behavior was found when the curvilinear trend was used. Further sensitivity analyses revealed no apparent influence of the level of program differentiation. In sum, our findings indicate that the GBG does not improve behavior when implemented in this sample of English schools.

Keywords

Good Behavior Game; externalizing behavior; differential effects; universal intervention; cumulative risk

Game over? No Main or Subgroup Effects of the Good Behavior Game in a Randomized Trial in English Primary Schools

Persistent low-level disruptive behavior (e.g. chatting, calling-out without permission) was highlighted as an area of concern by the Chief Inspector for schools in England (Office for Standards in Education [Ofsted], 2013). Such behavior is proposed to have a detrimental impact on life chances of students, and reduce retention among teachers (Ofsted, 2014). Teachers also share these concerns, with 69% of the 1048 surveyed by Ofsted identifying talking and chatting as a key problem in every lesson. Additional concerns in most lessons included calling-out without permission, failing to follow instructions, and fidgeting. Over 25% reported that the impact of this on learning was high (Ofsted, 2014). While these behaviors are not as severe as the aggressive and anti-social behaviors that characterize conduct disorders (National Institute for Health and Care Excellence, 2013), all behavior problems are likely to impact on the learning, participation and achievement of students, and it is estimated that up to an hour of learning is lost each day as a direct consequence of low-level disruption in classrooms (Ofsted, 2014).

Korpershoek and colleagues' recent meta-analysis (2016) produced a useful taxonomy of different approaches to classroom management, namely *teachers' behavior-focused*, *teacher-student relationship-focused*, *students' behavior-focused*, and *students' social-emotional development-focused* interventions. One such *students' behavior-focused* intervention is the Good Behavior Game (GBG; see Intervention section in Method), an interdependent group-contingency behavior management strategy (Lastrapes, 2013) that aims to target low-level disruptive behaviors that interfere with learning, in order to allow more time to teach (Chan, Foxcroft, Smurthwaite, Coomes, & Allen, 2012). It was originally developed by Barrish, Saunders, and Wolf (1969) in the United States of America (USA) and is designed to be used by teachers alongside the curriculum in elementary schools.

Evidence spanning several decades across many countries worldwide (e.g. Dion et al., 2011; Leflot, Van Lier, Onghena, & Colpin, 2010; Ruiz-Olivares, Pino, & Herruzo, 2010; van Lier, Muthén, van der Sar, & Crijnen, 2004) attests to the impact of the GBG on behavioral outcomes (Flower, Mckenna, Bunuan, Muething, & Vega, 2014). This research can be broadly categorized into three domains: low-level disruptive behaviors, conduct problems, and aggressive behaviors, with disruptive behavior the outcome most commonly examined (Flower et al., 2014). However, while other positive effects are outlined in the program logic model (Chan et al., 2012), less evidence exists regarding the effectiveness of the GBG in improving these outcomes. For instance, while increased on-task behavior is theorized as an immediate outcome of GBG implementation, few studies have examined this explicitly. One exception found that an adapted version of the GBG in Canada improved students' attentional focus (Dion et al., 2011). The program logic model also outlines increased pro-social behaviors and social awareness as immediate impacts of the GBG, in addition to increased positive peer interactions and reduced anti-social behavior as short- and medium- term impacts (Chan et al., 2012). Indeed, a pilot study of the GBG in Oxfordshire, England found that pro-social behavior improved, along with social competence and decreased social isolation. Qualitative data also supported this, with teachers reporting more effective interpersonal communication among students, and increased sociability (Coombes, Chan, Allen, & Foxcroft, 2016).

However, a recent meta-analysis (Flower et al., 2014) highlighted some inconsistencies in findings regarding the impact of the GBG on children's behavior. For example, while several studies have reported a positive impact of the GBG on disruptive behaviors (e.g., Barrish et al., 1969; Kleinman & Saigh, 2011; Saigh & Umar, 1983), Leflot and colleagues' (2010) study in Belgium found no significant impact of the program on out-of-seat behaviors. These null findings may be the result of adaptations that were made to the

GBG in the Dutch[1] version of the game. Indeed, Coombes and colleagues (2016) emphasize the importance of adhering to the manualized procedures specified by the program developers.

The GBG in England: Issues of Cultural Transferability and Program Differentiation

The inconsistent findings noted above may also be due to cultural incompatibility of the GBG, as the vast majority of the research has been conducted in the USA (Humphrey et al., 2016; Lendrum & Humphrey, 2012). Indeed, a recent meta-analysis of school-based interventions found that larger effect sizes for some outcomes are evidenced when a program is implemented in its country of origin (Wigelsworth et al., 2016). It is thought that both local needs and fit with the new cultural context are factors that can influence the success of these interventions (Castro, Barrera, & Martinez, 2004), and so some aspects of the English school system may impact on the delivery of the GBG. For example, the National Curriculum and priorities outlined by Ofsted affect the time and resources that teachers have to successfully implement optional programs. Furthermore, some teachers have previously noted the prohibition of teacher-student interaction during GBG gameplay sessions as problematic (Ashworth, Demkowicz, Lendrum, & Frearson, 2018; Chan et al., 2012). This may reduce the social validity of the intervention (e.g., its acceptability, feasibility and utility), which in turn is likely to influence the extent to which teachers in English schools adhere to the guidelines provided by the intervention developers, and thus the likelihood that implementation will be high and will be sustained (Wehby, Maggin, Partin, & Robertson, 2011).

Conversely, it is possible that the GBG may not be sufficiently distinct from teachers' usual practice to evidence significant gains in students' outcomes. In the decades since the intervention was first established, many of the procedures embodied within it have become standard behavior management practices (e.g., provision of rewards, classroom rules, group-based contingencies, monitoring behavior). This speaks to the concept of program

[1] The GBG was implemented in a Flemish-speaking (Dutch) area of Belgium

differentiation, defined as, “the extent to which a program’s theory and practices can be distinguished from other programs” (Durlak & DuPre, 2008, p. 329). This aspect of implementation has been sorely neglected in research (e.g., 0/59 studies documented in Durlak and DuPre’s (2008) seminal review), despite its potential importance as a moderator of outcomes. Low levels of program differentiation may be advantageous because a given intervention will feel more familiar to staff and can be assimilated more easily into existing processes and practices; conversely, high levels may be desirable in that the intervention will be seen as more distinctive, adding value to what is already in place (Humphrey, 2013). Either way, it is important to take into account the “uniqueness” of an intervention when evaluating its impact, in order to understand what led to any change in students’ outcomes (or lack thereof; Humphrey et al., 2016).

Differential Gains

The intention-to-treat (ITT) principle, whereby analyses include every subject who is randomized, regardless of anything that happens after randomization (e.g., noncompliance, protocol deviations, withdrawal), dominates analyses of randomized trials as it provides an unbiased effect estimate (Gupta, 2011). However, it is well established that students do not respond in a uniform manner to universal interventions; natural heterogeneity exists within student populations, and these interventions can differentially affect various strata of the population (Greenberg & Abenavoli, 2017). Indeed, students deemed to be “at-risk” typically evidence greater benefits (Farrell, Henry, & Bettencourt, 2013). For example, an evaluation of Second Step, a universal preventive intervention, found 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).

There has been a strong focus in GBG research on the benefits for certain “at-risk” groups, namely boys and those from low socio-economic backgrounds. Results have consistently shown that these groups of students, and in particular boys who were found to be highly aggressive at baseline, evidenced greater gains from the intervention (Dolan et al., 1993; Kellam et al., 2011; Kellam, Rebok, Ialongo, & Mayer, 1994; Kellam, Ling, Merisca, Brown, & Ialongo, 1998). Studies have also examined the effects of the intervention on individuals displaying "early risk behaviors" for substance abuse, depression and antisocial behavior, reporting positive results (e.g. Ialongo et al., 1999). Indeed, the emphasis on differential gains of the GBG is so strong, some studies *only* report subgroup effects (e.g., effects reported by gender; Dolan et al., 1993); although this brings into question the overall effectiveness of the intervention at the ITT level.

However, extant GBG research that utilizes the term "at-risk" generally uses this as a proxy for "highly aggressive" (Dolan et al., 1993). This implies that risk is binary, whereas evidence suggests that risk factors cluster together and are not independent of each other, as proposed in cumulative risk (CR) theory (Rutter, 1979). Studies that have adopted a CR model of risk when examining outcomes for children have typically found that the *number* of risk factors present is a superior predictor of negative outcomes than the *nature* of the individual risks (Evans, Li, & Whipple, 2013). It is thought that measuring the effects of risk factors in isolation – as has been the case in existing studies of the GBG – can over-estimate the importance of a given factor by not accounting for the complex relationships between them (Gerard & Buehler, 1999; Sameroff, Gutman, & Peck, 2003). It is the confluence of various risk factors, rather than any particular factor, that leads to dysfunction (Flouri & Kallis, 2007).

Thus, research which adopts a CR perspective represents a potentially important step forward in examining potential subgroup effects of preventive interventions, as it more accurately represents individual differences and the multitude of factors influencing a child's development. However, to the authors' knowledge, only one study to date has utilized a CR approach when determining the effectiveness of a school-based intervention. An evaluation of the "Guiding Responsibility and Expectations in Adolescents Today and Tomorrow" (GREAT) student curriculum (The Multisite Violence Prevention Project, 2008) found that short- and long-term effects 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. The authors argued that this differential pattern of intervention effects may explain why main effects are not typically found in evaluations of universal interventions in middle schools, thus highlighting the importance of looking beyond the ITT approach.

It is theorized that the variation in outcomes of universal interventions for different subgroups is due to the extent to which the individuals within them display deficiencies in the skills targeted by the intervention (Farrell, Henry, & Bettencourt, 2013; Greenberg & Abenavoli, 2017). It therefore follows that those children at the higher levels of risk for the intended outcomes of the intervention, and are thus the most in need of it, will evidence the greatest gains. To date, however, "no research has yet examined the effectiveness of the GBG in relation to baseline risk profiles reflecting different constellations of risks across developmental domains" (2013, p. 480), although there is conjecture. For example, Muthén and colleagues (2002) hypothesized: "GBG may have its largest effect for those who are in the middle trajectory class, showing milder forms of problems, while not being strong enough to affect the most seriously aggressive children and not needed for members of the stable non-aggressive group" (p.461).

The Current Study

The aforementioned pilot study of the GBG conducted in six English primary schools for one year found positive effects on various aspects of students' behavior (Chan et al., 2012;

Coombes et al., 2016). However, this study did not utilize a control group, thus limiting the extent to which these improvements could be securely attributed to the intervention. Therefore, the current study, a cluster-randomized controlled trial (RCT) of the GBG in England, intended to extend this previous research by conducting a more rigorous evaluation utilizing a larger, more representative sample. In addition to examining the impact of the GBG on student behavior (namely disruptive behavior, concentration problems, and pro-social behavior) in English schools, a sensitivity analysis was also conducted to establish whether students' behavioral outcomes varied as a function of levels of program differentiation. Furthermore, pre-specified subgroup analyses were conducted to explore any potential differential effects of the intervention on children at low, medium and high levels of CR exposure. To determine the sensitivity of our findings to changes in the modeling of children's risk status (Evans et al., 2013), additional analyses were conducted, whereby CR exposure (both the linear and quadratic terms, the latter of which represents the curvilinear trend of the continuous CR measure) was treated as a continuous variable.

Thus, our research questions were as follows:

1. Do children in primary schools implementing the GBG over a two-year period demonstrate significant improvements in behavior (specifically, a) disruptive behavior, b) concentration problems and c) pro-social behavior) compared to those children attending usual practice schools?
2. Are any findings in relation to RQ1 sensitive to varying levels of program differentiation?
3. Are there differential intervention gains in behavioral outcomes among children at different levels of cumulative risk exposure?
4. Are any findings in relation to RQ3 sensitive to changes in the way risk exposure is modeled?

Method

Design

The current study utilizes data from an efficacy trial of the GBG in England and so methods have been published previously. In brief, a two-year cluster-randomized design was utilized, with participating schools as the unit of randomization. A local trials unit randomly allocated schools to one of two trial arms: (1) GBG (intervention arm); or (2) usual provision (UP arm). To ensure balance across the arms of the trial, a minimization algorithm (adaptive stratification) was applied regarding the proportion of children eligible for free school meals (FSM) and school size. 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 [masked for peer review].

Schools were recruited between March and July 2015. Eligible schools were mainstream, state-maintained primary schools (serving children aged four-11 years). Participation required informed consent from the schools' Head Teachers. Child assent and parental opt-out consent were also sought. In total, 68 parents (2.2%) exercised their right to opt their children out of the trial, and no children declined assent or exercised their right to withdraw from the study. The study received approval from the ethics committee of the authors' host institution.

Participants

The trial sample were N=3084 children aged six-seven in 77 schools in three regions across England (see supplemental files for CONSORT diagram). The composition of participating schools mirrored that of primary schools in England regarding size and the proportion of students speaking English as an Additional Language (EAL). However, trial schools typically had significantly larger proportions of children with special educational needs and disabilities (SEND) and those eligible for FSM, in addition to lower rates of absence and attainment. At the student level, the trial sample were also generally above the national average regarding the proportion who were identified as having an SEND, eligible for FSM, and speaking EAL; while they were typically below average with regards to attainment (DfE, 2015; Table 1). Differences between schools and students across the trial arms were negligible, indicating good balance and successful randomization.

Sample Size

Sample size calculations were carried out using Optimal Design Software at the point of randomization. With an intra-cluster correlation co-efficient (ICC) of 0.08 for the primary outcome measure (disruptive behavior) at baseline, a pre-post correlation of 0.63, an average cluster size of 40, and standard Power and Alpha thresholds of 0.80 and 0.05 respectively, the minimum detectable effect size (MDES) for an ITT analysis was determined to be 0.16. Given this, the trial was considered to be well powered.

Intervention

The GBG is an "interdependent group-oriented contingency management procedure" (Tingstrom, Sterling-Turner, & Wilczynski, 2006, p. 225) designed to be integrated into the existing curriculum without taking up any additional teaching time. It is underpinned by behaviorism (i.e., reinforcement of desired behaviors; Skinner, 1948), social learning theory (i.e., vicarious learning through the modeling of appropriate behaviors; Bandura, 1977), and life course/social field theory (i.e., an individuals' ability to meet the social demands of a particular environment; Kellam, Branch, Agrawal, & Ensminger, 1975). Core components are (1) *classroom rules,* (2) *team membership,* (3) *monitoring behavior*, and (4) *positive reinforcement*. It is suggested that the GBG should initially be implemented three times a week, for ten minutes each time, with this increasing to everyday for up to 30 minutes over the course of the year. It should also be played at varying points in the day, during an assortment of lessons and activities. The logic model for the program is available in Chan et al. (2012).

When playing the game, students are divided into teams of up to seven that are gender-balanced and heterogeneous in behavior and academic ability. Teams are expected to follow four rules during the game: (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); teachers monitor behavior and record any infractions that occur as a result of a team member failing to follow these. Other than when recording an infraction, teachers should not interact with students during the game. In order to win the GBG, and thus access agreed rewards or privileges, teams need to have four or fewer infractions at the end of the game. 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).

As the trial took place over a two-year period, different teachers delivered the GBG in the first and second years. All teachers attended two days of training prior to implementation, in the September or October of their delivery year, with a further day of top-up training a few months later. Training focused on the theoretical underpinnings of the GBG and procedures for implementation. Trained GBG coaches visited schools approximately once per month throughout the trial to support teachers' implementation (e.g. observation and feedback, modeling delivery; Ashworth et al., 2018). Coaches had all previously worked as teachers or education professionals and were trained by the program developers.

Implementation

Implementation fidelity/quality, participant responsiveness and reach were assessed via annual structured observations in the second term of each year of the trial (January-April). These were developed and piloted using video footage of GBG delivery in English schools recorded in the aforementioned UK pilot (Chan et al., 2012). Inter-rater reliability was found

[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.

to be very high. Dosage was measured via an online scoreboard developed for teachers to record infractions during the game; this also included details of length and frequency of play.

Average scores for fidelity/quality (2015/16: 69.79%; 2016/17: 70.11%) were high, suggesting that teachers followed most of the prescribed steps in the manual and did so in an enthusiastic and engaging manner. Participant responsiveness was also high (2015/16: 74.51%; 2016/17: 69.07%), as was participant reach (2015/16: 95.26%; 2016/17: 95.98%), indicating that students responded favorably to the GBG and were largely present for delivery. The intervention was implemented on average twice per week in the first year of the trial (2015/16: 1.93 games per week), although this reduced slightly in the second year (2016/17: 1.55 games per week); the average game lasted approximately 15 minutes in both years. Thus, while average *duration* was well within the range of the only other GBG trials that have reported dosage data, the *frequency* of game play was somewhat lower (Domitrovich et al., 2015; Hagermoser Sanetti & Fallon, 2011; Kellam et al., 1998; Pas et al., 2015).

Measures

Behavior. Behavior was assessed using the checklist version of the Teacher Observation of Classroom Adaptation (TOCA-C; Koth, Bradshaw, & Leaf, 2009). This 21-item scale assesses students' concentration problems (inattentive and off-task behavior; seven items), disruptive behavior (disobedient, disruptive and aggressive behaviors; nine items) and pro-social behavior (positive social interactions; five items). Statements are provided about the child (e.g. gets angry when provoked by other children), which teachers read and endorse on a six-point scale (Never/Rarely/Sometimes/Often/Very Often/Almost Always). Children's scores are then summed and averaged (one-six), with higher scores indicating more maladaptive behaviors for concentration and disruptive behavior, and lower scores indicative of poorer pro-social behavior (Kourkounasiou & Skordilis, 2014).

The TOCA has been frequently used in previous research on the GBG (e.g. Bradshaw et al., 2015; Chan et al., 2012; Kellam et al., 1994). The checklist version has good psychometric properties, including high internal consistency (all subscales α>0.86), and a factor structure that is invariant across gender, race and age (Koth et al., 2009; Bradshaw, Waasdorp, & Leaf, 2015). Internal consistency of the TOCA-C subscales in the trial was excellent (all α>0.87 at baseline).

Program differentiation. As a means through which to determine the level of program differentiation of the GBG, usual practice surveys (based on existing measures of classroom management strategies; Reupert & Woodcock, 2010) were administered to all teachers at baseline. Twelve items deemed to reflect the presence or absence of key GBG procedures and practices (e.g. establishing and maintaining a set of classroom rules; observing and monitoring students' behavior in the classroom; use of prizes as rewards for good behavior; use of group rewards; use of a warning/strike system) were extracted to create a program differentiation index (PDI), with scores ranging from 0-12; higher scores were indicative of lower levels of program differentiation. PDI scores were transformed to percentages for ease of interpretation and aggregated to the school-level. Finally, the school-level PDI score was converted to a binary variable, with schools categorized into either low or moderate[3] program differentiation (utilizing the 50th percentile as a cut-point). 38 schools (17 GBG; 21 UP) were designated as moderate PDI and 38 (21 GBG; 17 UP) as low PDI. The remaining school's PDI score could not be calculated as they failed to complete usual practice surveys.

Cumulative risk. Previous research by the authors utilizing the same dataset identified the student- and school-level risk factors that were significant predictors of baseline disruptive behavior scores. These analyses were subsequently extended for the present study to incorporate concentration problems and pro-social behavior (see Table 2). As CR theory

[3] Given the distribution of scores, which ranged from 56-92%, 'moderate' and 'low PDI was deemed to be more accurate than 'high' and 'low'.

states that the number of risk factors is more important than their nature, these risk factors were then dichotomized (coded as either '0' for absent or '1' for present) and summed for each of the behavioral outcomes, creating three CR scores for each child that represented the number of risk factors to which they were exposed. This is consistent with previous work in the field (e.g., Ashworth & Humphrey, 2018; Gerard & Buehler, 2004; Hebron, Oldfield, & Humphrey, 2016; Oldfield, Humphrey, & Hebron, 2015). Calculations of effect size were conducted to determine where differences in mean behavior scores between risk levels lay; line graphs were also plotted to provide a visual representation of the risk-outcome relationship (see supplemental files). Risk groups were then created in accordance with the elbow points present on these graphs, and where differences between risk levels were notable. Children were therefore categorized into one of three groups for each of the three measures of behavior: *low-risk*, *medium-risk*, or *high-risk* (see Table 2).

Analysis

ITT analyses (controlling for school-level FSM and school size, and child-level gender, FSM and SEND status) were conducted and subsequently extended to incorporate an analysis of subgroup moderator effects. Multi-level modeling (in MLwiN 2.36) was used to account for the clustered and hierarchical nature of the data (students nested within schools; Twisk, 2006). A fixed effects random intercepts model was utilized, which assumes that baseline scores will have different explanatory power for different schools, and that the relationship between baseline and follow-up will not vary by school. Prior to analysis, behavior scores were standardized by converting them to z scores, meaning that the coefficients reported can be interpreted as effect sizes akin to *Cohen's d*, thus facilitating interpretation across models (Bierman et al., 2014).

First, two-level models were fitted (one for each measure of behavior) with intervention group allocation at the school-level (with *UP allocation* utilized as the reference category) and the relevant baseline behavior scores and student-level covariates (gender, SEND, FSM) at the student-level as explanatory variables, to establish any main effects of the GBG on the outcome variables (post-test behavior scores). Second, PDI group was added at the school-level as an explanatory variable, and interaction terms between trial group and PDI (with *low PDI* utilized as the reference category) were specified using dummy coding (e.g. 0 = low, 1 = moderate) to establish whether our substantive findings were sensitive to levels of program differentiation.

For research question 2, risk group categorization was added to the original model at the student-level. Cross-level interaction terms between the intervention group and the three risk groups (with the *no risk* group utilized as the reference category) were specified using dummy coding, to establish any subgroup moderator effects. Finally, to determine the sensitivity of our substantive findings to changes to the modeling of children's risk status, two additional multi-level models were fitted. The original form of the continuous CR score was added to the first, along with an interaction term between CR score and intervention group. For the second, consistent with previous literature (Ashworth & Humphrey, 2018; Oldfield et al., 2015), the continuous CR score was mean-centered and squared, to generate a quadratic CR score, and was then added to the model, along with an interaction term between quadratic CR score and intervention group.

18.5% of participants in the sample had incomplete data, in cases where they had left the school (12.6%) or teachers had failed to provide post-test behavior data (5.9%). Missing value analysis was conducted through binary logistic regression to identify the variables that predicted partially observed data. Missingness was predicted by school size, school-level absence, school-level behaviour, student SEND status, and student looked-after status. Thus, data were likely to be missing at random. Therefore, in order to maintain the sample size, multiple imputation (MI) procedures were implemented. This reduces the bias associated with attrition and allows for the use of statistical techniques designed for complete datasets (Pampaka, Hutcheson, & Williams, 2016). MI was conducted in REALCOM-Impute with demographic variables and trial group allocation added as auxiliary (where data were fully observed) and response variables. REALCOM-Impute default settings of 10 datasets, 1000 iterations, a burn-in of 100, and a refresh of 10 were utilized, in accordance with guidance produced by Carpenter and colleagues (2011) for multi-level imputation with mixed response types.

Results

Descriptive statistics pertaining to both the main and subgroup analyses are presented in Table 3.

Research Question 1

ITT analyses (Table 4) indicated that the GBG had no overall effect on children's a) disruptive behavior (d = 0.056, p = .235), b) concentration problems (d = 0.022, p = .400) or c) pro-social behavior (d = -0.108, p = .160).

Research Question 2

There was no significant interaction found between moderate levels of program differentiation and GBG trial group allocation for a) disruptive behavior (d = 0.019, p = .451), b) concentration problems (d = 0.053, p = .378), or c) pro-social behavior (d = -0.282, p = .097) (see Table 5).

Research Question 3

There were no statistically significant subgroup effects of the GBG on a) disruptive behavior (medium-risk: d = 0.055, p = .217; high-risk: d = -0.234, p = .100), b) concentration problems (medium-risk: d = -0.086, p = .137; high-risk: d = -0.098, p = .179), or c) pro-social behavior (medium-risk: d = -0.027, p = .362; high-risk: d = 0.191, p = .165) for students in any of the risk groups, relative to the low-risk group (Table 6).

Research Question 4

The sensitivity analysis (Table 7), whereby CR was modeled as a linear continuous variable, confirmed previous findings, with no significant interaction found between trial group and risk level in predicting post-test a) disruptive behavior (*d* = 0.004, p = .453), b) concentration problems (*d* = -0.048, p = .057), or c) pro-social behavior (*d* = 0.034, p = .187). With regards to the quadratic term (Table 8), a small but significant interaction was found between trial group and risk level in predicting post-test disruptive behavior only (*d* = -0.092, p = <.001). No such effects were found for concentration problems (*d* = 0.002, p = .462) or pro-social behavior (*d* = 0.020, p = .213).

Discussion

The results of this RCT demonstrate that the GBG had no main effect on students' disruptive behavior, concentration problems, or pro-social behavior. Allocation to the intervention group was also not a statistically significant predictor of outcomes for students exposed to varying levels of CR. In other words, exposure to the GBG for two years did not result in significant improvements in students' behavior, irrespective of risk status. These findings were insensitive to levels of program differentiation, and were largely unaffected by the way that risk exposure was modeled, although a significant effect was found for disruptive behavior when CR exposure was modeled using the quadratic term. However, given the large number of comparisons conducted, the latter finding may well be due to familywise error as opposed to a genuine effect. Furthermore, the size of the subgroup moderator effect was very small, and thus this finding was not considered to be practically meaningful.

No Main Effects of the GBG on Behavior

As noted previously, studies of the GBG typically find positive effects for students' behavior; therefore the findings from the present study are incongruent with much of the existing literature in the field. However, of the three measures of behavior utilized in the present study, only disruptive behavior has previously been examined in-depth. Although the program logic model also outlines improvements in on-task behavior and pro-social behavior as immediate impacts of the GBG (Chan et al., 2012), Flower and colleagues' meta-analysis (2014) found that only two studies had tested the effects of the intervention on social outcomes (antisocial behaviors and social interactions), and none had examined concentration explicitly (only attention). Thus, while these outcomes are hypothesized to occur as a result of the GBG by the program developers, they have not been extensively tested. Furthermore, studies of the GBG often find greater subgroup than main effects on behavior, namely for boys and students from low socio-economic backgrounds, and some studies only report subgroup effects (e.g. Bradshaw, Zmuda, Kellam, & Ialongo, 2009; Dolan et al., 1993). Thus, when looking specifically at ITT analyses, and more infrequently tested outcomes, it was unclear if any main effects would be found.

Alternatively, the null results in the present study may be explained by a lack of cultural transferability of the GBG. Although previous research has found the intervention to be effective outside of its country of origin, it was adapted prior to implementation to suit the culture of the countries in which the studies were located (e.g. France, Spain, the Netherlands; Dion et al., 2011; Ruiz-Olivares et al., 2010; van Lier et al., 2004). However, in the present study, the GBG was implemented in its original format. Thematic analyses of the qualitative data from both the Oxfordshire pilot (Chan et al., 2012) and the IPE associated with the current study indicated that teachers had concerns regarding several aspects of the intervention, namely the lack of teacher-student interaction permitted, and the inflexibility of the program not allowing teachers to adapt it to suit their classes' needs. Thus, it is possible that the intervention was not compatible with the school culture in England, and hence its effects were diluted (Wigelsworth et al., 2016).

It is therefore also possible that the null results were due to implementation failure, whereby teachers failed to adhere to the steps outlined in the manual due to perceived incompatibility of the intervention. Indeed, almost one-quarter of schools ceased implementation over the course of the trial (although they still complied with data collection protocols, and there were no significant differences between teachers who ceased and sustained implementation). Furthermore, dosage (frequency and duration of game play) was also lower than is recommended (Ford, Keegan, Poduska, Kellam, & Littman, 2014), meaning that the game was not implemented as often or for as long as is specified by the program developers. Thus, it may be that school timetables and other demands placed on teachers (Education Committee, 2017) mean that the GBG does not fit well in English schools. However, the other studies of the GBG in the USA that have measured dosage have reported similar issues (e.g. Domitrovich et al., 2015; Hagermoser Sanetti & Fallon, 2011), meaning that the time to implement the GBG is unlikely to be an issue specific to the English schooling system. Furthermore, while a certain level of dosage is specified in the intervention’s manual, these dosage benchmarks have not been empirically validated, and so the levels of dosage necessary to bring about student gains are unknown (Becker, Darney, & Domitrovich, 2013).

As the present study was an efficacy trial, in which significant resources were made available to optimize implementation (e.g. developer support, subsidized costs), the levels of implementation reported here are likely the ‘best case’ scenario (that is, we could reasonably expect further dilution of implementation under ‘real world’ conditions). Furthermore, procedural fidelity, quality, participant responsiveness and reach recorded in our structured observations were high, indicating that with the exception of suboptimal dosage, implementation failure was likely not a key issue. Nevertheless, it is important that future studies examine whether null results are still evident once implementation variability is accounted for using appropriately robust methods (e.g. complier average causal effect estimation; CACE). Indeed, a CACE analysis would not only help to determine whether the null results found in the present study were a result of suboptimal dosage, but would also remove the potential bias introduced in ITT (wherein full compliance with the intervention is assumed) analyses (Peugh, Strotman, McGrady, Rausch, & Kashikar-Zuck, 2017). This will be formally examined in a future paper.

Alternatively, it is possible that the GBG was not required in English schools. First, behavior management strategies akin to key aspects of the GBG were commonplace in classrooms at baseline, and so it may not have been distinct enough to bring about additional gains in students' behavior. Thus, it may not be that the GBG was ineffective, but that it was simply no *more* effective than the existing behavior management strategies already in use both at baseline and in the usual provision group during the trial. Our sensitivity analysis demonstrated that the interaction between trial group allocation and program differentiation level did not predict post-test outcomes, indicating that the intervention was equally unsuccessful in improving behavior in contexts in which it was more distinct from existing practice. However, it should be borne in mind that, such was the proliferation of GBG-like strategies in use among all participating teachers, it was not possible to model outcomes in the context of high program differentiation. Furthermore, as the collection of usual practice data is not standard procedure when evaluating the effectiveness of an intervention, this finding cannot be compared to previous studies of the GBG that have identified a positive effect. This therefore highlights the importance of collecting this type of data when conducting an RCT.

Second, contrary to widely publicized concerns noted at the outset of the current study, other data challenges the deficit view of student behavior in English schools. For instance, Ofsted's most recent report identified that 92.3% of all schools in England were judged Good or Outstanding for behavior standards (DfE, 2012). Indeed, students' average behavior scores on the TOCA-C were low at baseline across the trial arms (e.g., disruptive behavior = 1.71/6), leaving little room for improvement.

No Subgroup Effects Among Students Exposed to Varying Levels of Cumulative Risk

The current study was among the first to apply CR theory to subgroup moderator analyses of behavior in a preventive intervention trial. Although no subgroup effects were found in the present study for students at any risk level, the analyses in the wider trial using 'traditional' subgroups (boys at-risk of conduct problems and children eligible for FSM) also found no evidence of statistically significant differential gains. This contradicts the majority of GBG studies, which typically find such effects for students using a single risk factor marker (e.g. Dolan et al., 1993; Kellam et al., 2011; Kellam, Ling, Merisca, Brown, & Ialongo, 1998; Kellam, Rebok, Ialongo, & Mayer, 1994). Similarly to the above discussion, it is possible that cultural incompatibility or implementation failure provide an explanation for these incongruous results. Thus, the null findings were likely not a reflection of the choice to utilize CR theory to measure subgroup effects, but instead suggest that the GBG simply did not have any impact on at-risk children, regardless of the approach taken to the modeling of risk.

This, therefore, does not rule out the utility of CR indices when examining subgroup effects of preventive interventions. Indeed, there is evidence from other studies to suggest its utility (The Multisite Violence Prevention Project, 2008). In line with ecological systems theory (Bronfenbrenner, 1986), it is thought that CR theory provides a more accurate representation of a child's experiences regarding risk exposure, accounting for the clustering of risk factors and interactions between them that are likely to occur (Flouri & Kallis, 2007). However, as the subgroup moderator analysis was exploratory in nature, the small sample sizes of the different risk groups mean that the study may have been under-powered. Thus, larger-scale explorations of subgroup effects utilizing CR indices warrant further attention in future research.

Implications

The results from this study highlight the importance of reporting null results of trials (Fiennes, 2018), helping to reduce the disconnect between scientific worth and culture (Matosin, Frank, Engel, Lum, & Newell, 2014). In addition to tackling the widely publicized publication bias (which favors statistically significant, 'positive' results; Fanelli, 2010), these null results also advance knowledge in the field and are important for several stakeholders. For instance, this finding is beneficial to both funding bodies and schools that need to be aware of the utility of an intervention in order to make an informed decision to implement it with students. This is particularly important when a wealth of evidence already exists regarding the intervention's efficacy in its country of origin, but it has not previously been rigorously tested within the local culture, as decreases in effectiveness are often identified once an intervention is exported (Wigelsworth et al., 2016). Furthermore, it is vital that intervention developers know if their program is ineffective outside of its country of origin; this information allows them to make decisions regarding necessary adaptations to ensure the intervention's viability in different countries. Indeed, the finding that the GBG is not effective in achieving its primary intended outcome is a key issue that needs to be addressed.

The findings from the present study also highlight the broader need to evaluate the cultural transferability of previously successful interventions before they are exported and implemented on a large scale. Significant adaptations may be necessary if the expected outcomes are to be achieved, as they can enhance ownership and commitment, support 'goodness-of-fit' to the local culture, and improve sustainability (Lendrum & Humphrey, 2012). Indeed, previous research suggests that a major factor in the successful transferability of interventions is their adaptability (Castro et al., 2004). Thus, future research should seek to identify the cultural adaptations of both the GBG and other school-based interventions that ensure that they are suitable to the English context, while also ensuring the program's critical components are still in place (Sharples, Albers, & Fraser, 2018).

Strengths and Limitations

There are several factors that increase the security of the findings noted above. The use of a cluster-randomized design with well-balanced trial arms at the school- and student-levels means that the likelihood of diffusion or contamination effects was minimized (Campbell, Mollison, & Grimshaw, 2001). In addition, the trial was well powered to detect effects (MDES of 0.16), and although attrition was at 18.5% at the student-level, this was within acceptable limits (Dumville, Torgerson, & Hewitt, 2006) and was appropriately addressed using MI. The use of sensitivity analyses to establish that our substantive findings did not vary by program differentiation levels or the way in which CR status was modeled further increases the security of these results.

Nevertheless, it is possible that null results were the result of research design limitations. For instance, while all schools in the desired regions were invited to participate in the trial, the schools in the sample were typically larger than average, with higher rates of students with an SEND, eligible for FSM, and speaking EAL (DfE, 2015). Indeed, the majority of schools were situated in one densely populated, ethnically diverse region with high levels of socio-economic deprivation. In addition, schools that chose to participate in the trial were likely those that had a greater perceived need for a behavior management intervention. As such, the schools participating in the trial may not have been fully representative of those in England overall. Furthermore, as class composition was not known for schools in the UP arm of the trial, it was not possible to include teachers as a level of analysis in the model. Thus, we therefore missed the opportunity to model teacher characteristics as a potentially strong source of variance.

While our primary outcome measure, the TOCA-C, has previously been validated, psychometric validation studies are limited, and have typically been conducted by developers of either the TOCA or GBG (e.g. Bradshaw et al., 2015; Kellam et al., 1994). As the TOCA was designed by a developer of the GBG, and has primarily been used in GBG studies, there may be concerns that the measure is ‘inherent to treatment’ (Slavin & Madden, 2011). However, as this issue would have *favored* positive outcomes in the intervention arm of the trial, and no effects of the GBG were identified, it is unlikely that this was an issue. In addition, the TOCA-C is a teacher informant-report measure. Its developers advise that a variety of factors (e.g. the demographic characteristics of the child) can influence teachers’ reports of behavior problems. It is also thought that the timing of administration can influence reports, with a notable difference in scores between the beginning and end of the school year (Dolan et al., 1993; Koth et al., 2009). However, such issues apply equally to both trial arms and are therefore unlikely to have biased our findings. Finally, it is acknowledged that no single informant can provide a comprehensive picture of a student’s behavior. Inter-rater correlations between teacher- and self-report behavior measures have previously been found to be weak (Goodman, Meltzer, & Bailey, 1998) and so collecting similar additional data from other informants may have provided a more comprehensive and valid assessment (De Los Reyes et al., 2015).

Conclusions

The present study is, to the authors’ knowledge, the largest RCT of the GBG worldwide to date, and is the first in an English setting, thus providing a significant contribution regarding the efficacy of the intervention and its cultural transferability when delivered in its original format. The current study also contributes to the evidence base regarding the effects of the GBG on concentration problems and pro-social behavior, examining the previously infrequently tested claims outlined in the program logic model.

Furthermore, the study has advanced the examination of differential gains by testing subgroup effects for students at varying levels of CR exposure, an area highlighted as a priority for future research in Durlak and colleagues' (2011) meta-analysis, and the influence of program differentiation, an aspect of implementation sorely neglected in prior literature (Durlak & DuPre, 2008).

Acknowledgements

This work was supported by funding from [masked for peer review].

Declaration of Interest Statement

The authors declare they have no conflicts of interest.

References

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*, *10*. https://doi.org/10.1007/s12310-018-9282-3

Bandura, A. (1977). *Social learning theory*. Englewood Cliffs, NJ: Prentice Hall.

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

Becker, K., Darney, D., & Domitrovich, C. (2013). Supporting universal prevention programs: A two-phased coaching model. *Clinical Child and Family Psychology Review*, *16*, 213–228. https://doi.org/10.1007/s10567-013-0134-2.Supporting

Bierman, K., Nix, R., Heinrichs, B., Domitrovich, C., Gest, S., Welsh, J., & Gill, S. (2014). Effects of Head Start REDI on children's outcomes 1 year later in different kindergarten contexts. *Child Development*, *85*, 140–159. https://doi.org/10.1111/cdev.12117

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

Bradshaw, C., Zmuda, J., Kellam, S., & Ialongo, N. (2009). Longitudinal impact of two universal preventive interventions in first grade on educational outcomes in high school. *Journal of Educational Psychology*, *101*, 926–937. https://doi.org/10.1037/a0016586

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

Campbell, M., Mollison, J., & Grimshaw, J. (2001). Cluster trails in implemenation resarch: Estimation of intracluster correlation coefficients and sample size. *Statistics in Medicine*, *20*, 391–399.

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

Castro, F., Barrera, M., & Martinez, C. (2004). The cultural adaptation of prevention interventions: Resolving tensions between fit and fidelity. *Prevention Science*, *5*, 41–45.

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*.

Coombes, L., Chan, G., Allen, D., & Foxcroft, D. (2016). Mixed-methods evaluation of the Good Behaviour Game in English primary schools. *Journal of Community & Applied Social Psychology*, *26*, 369–387. https://doi.org/10.1002/casp

De Los Reyes, A., Augenstein, T., Wang, M., Thomas, S., Drabkci, D., Burgers, D., & Rabinowitz, J. (2015). The validity of the multi-informant approach to assessing child and adolescent mental health. *Psychological Bulletin*, *14*, 858–900. https://doi.org/10.1158/1940-6207.CAPR-14-0359.Nrf2-dependent

DfE. (2012). *Pupil behaviour in schools in England*.

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

Dion, E., Roux, C., Landry, D., Fuchs, D., Wehby, J., & Dupéré, V. (2011). Improving attention and preventing reading difficulties among low-income first-graders: A randomized study. *Prevention Science*, *12*, 70–79. https://doi.org/10.1007/s11121-010-0182-5

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

Dumville, J., Torgerson, D., & Hewitt, C. (2006). Reporting attrition in randomised controlled trials. *BMJ*, *332*, 969–971. https://doi.org/10.1136/bmj.332.7547.969

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

Education Committee. (2017). *Recruitment and retention of teachers.* London, UK.

Evans, G., Li, D., & Whipple, S. (2013). Cumulative risk and child development. *Psychological Bulletin*, *139*, 1342–1396. https://doi.org/10.1037/a0031808

Fanelli, D. (2010). Do pressures to publish increase scientists' bias? An empirical support from US states data. *PLoS ONE*, *5*. https://doi.org/10.1371/journal.pone.0010271

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

Fiennes, C. (2018). Charity begins with admitting we got it wrong. Retrieved from https://www.ft.com/content/49e715b6-8458-11e8-9199-c2a4754b5a0e

Flouri, E., & Kallis, C. (2007). Adverse life events and psychopathology and prosocial behavior in late adolescence: Testing the timing, specificity, accumulation, gradient, and moderation of contextual risk. *Journal of the American Academy of Child and Adolescent Psychiatry*, *46*, 1651–1659. https://doi.org/10.1097/chi.0b013e318156a81a

Flower, A., Mckenna, J., Bunuan, R., Muething, C., & Vega, R. (2014). Effects of the Good Behavior Game on challenging behaviors in school settings. *Review of Educational Research*, *84*, 546–571. https://doi.org/10.3102/0034654314536781

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

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.

Gerard, J., & Buehler, C. (2004). Cumulative environmental risk and youth maladjustment: The role of youth attributes. *Child Development*, *75*, 1832–1849. https://doi.org/10.1111/j.1467-8624.2004.00820.x

Goodman, R., Meltzer, H., & Bailey, V. (1998). The strengths and difficulties questionnaire: A pilot study on the validity of the self-report version. *European Child & Adolescent Psychiatry*, *7*, 125–130. https://doi.org/10.1080/0954026021000046137

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

Gupta, S. (2011). Intention-to-treat concept: A Review. *Perspectives in Clinical Research*, *2*, 109–112.

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

Hebron, J., Oldfield, J., & Humphrey, N. (2016). Cumulative risk effects in the bullying of children and young people with autism spectrum conditions. *Autism*, *21*, 291–300. https://doi.org/10.1177/1362361316636761

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. (2013). *Social and emotional learning: A critical appraisal*. London, UK: Sage.

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

Kellam, S., Branch, J., Agrawal, K., & Ensminger, M. (1975). *Mental health and going to* *school: The woodlawn program of assessment, early intervention, and evaluation*. Chicago: University of Chicago Press.

Kellam, S., Ling, X., Merisca, R., Brown, C. H., & Ialongo, N. (1998). The effect of the level of aggression in the first grade classroom on the course and malleability of aggressive behavior into middle school. *Development and Psychopathology*, *10*, 169–170. https://doi.org/10.1017/S0954579498001564

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.

Kellam, S., Mayer, L., Rebok, G., & Hawkins, W. (1998). Effects of improving achievement on aggressive behavior and of improving aggressive behavior on achievement through two preventive interventions: An investigation of causal paths. In B. P. Dohrenwend (Ed.), *Adversity, stress, and psychopathology* (pp. 591–600). New York, NY: Oxford University Press. https://doi.org/10.1038/nrm3860.RNAi

Kellam, S., Rebok, G., Ialongo, N., & Mayer, L. (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*, *35*, 259–281.

Kleinman, K., & Saigh, P. (2011). The effects of the Good Behavior Game on the conduct of regular education New York city high school students. *Behavior Modification*, *35*, 95–105. https://doi.org/10.1177/0145445510392213

Korpershoek, H., Harms, T., de Boer, H., van Kuijk, M., & Doolaard, S. (2016). A meta-analysis of the effects of classroom management strategies and classroom management programs on students' academic, behavioral, emotional, and motivational outcomes. *Review of Educational Research*, *86*, 643–680.

https://doi.org/10.3102/0034654315626799

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

Leflot, G., Van Lier, P., Onghena, P., & Colpin, H. (2010). The role of teacher behavior management in the development of disruptive behaviors: An intervention study with the Good Behavior Game. *Journal of Abnormal Child Psychology*, *38*, 869–882. https://doi.org/10.1007/s10802-010-9411-4

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

Matosin, N., Frank, E., Engel, M., Lum, J., & Newell, K. (2014). Negativity towards negative results: A discussion of the disconnect between scientific worth and scientific culture. *Disease Models & Mechanisms*, *7*, 171–173. https://doi.org/10.1242/dmm.015123

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.

National Institute for Health and Care Excellence. (2013). *Antisocial behaviour and conduct disorders in children and young people: recognition, intervention and management*. London: NICE.

Office for Standards in Education. (2013). *The report of Her Majesty's Chief Inspector of Education, Children's Services and Skills*. London: Ofsted.

Office for Standards in Education. (2014). *Below the radar: Low-level disruption in the country's classrooms*. London: Ofsted.

Oldfield, J., Humphrey, N., & Hebron, J. (2015). Cumulative risk effects for the development of behaviour difficulties in children and adolescents with special educational needs and disabilities. *Research in Developmental Disabilities*, *41*, 66–75. https://doi.org/10.1016/j.ridd.2015.05.010

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., Becker, K., Domitrovich, C., Berg, J., Musci, R., & Ialongo, N. (2015). Identifying patterns of coaching to support the implementation of the good behavior game: The role of teacher characteristics. *School Mental Health*, *7*, 61–73. https://doi.org/10.1007/s12310-015-9145-0

Peugh, J., Strotman, D., McGrady, M., Rausch, J., & Kashikar-Zuck, S. (2017). Beyond intent to treat (ITT): A complier average causal effect (CACE) estimation primer. *Journal of School Psychology*, *60*, 7–24. https://doi.org/10.1016/j.jsp.2015.12.006

Reupert, A., & Woodcock, S. (2010). Success and near misses: Pre-service teachers' use, confidence and success in various classroom management strategies. *Teaching and Teacher Education*, *26*, 1261–1268. https://doi.org/10.1016/j.tate.2010.03.003

Ruiz-Olivares, R., Pino, M., & Herruzo, J. (2010). Reduction of disruptive behaviors using an intervention based on the Good Behavior Game and the Say-Do-Report correspondence. *Psychology in the Schools*, *47*, 1046–1058. https://doi.org/10.1002/pits.20523

Rutter, M. (1979). Maternal deprivation, 1972-1978: New findings, new concepts, new approaches. *Child Development*, *50*, 283–305.

Saigh, P., & Umar, A. (1983). The effects of a Good Behavior Game on the disruptive behavior of Sudanese elementary school students. *Journal of Applied Behavior Analysis*, *16*, 339–344.

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

Sharples, J., Albers, B., & Fraser, S. (2018). *Putting evidence to work: A school's guide to implementation*. London, UK: EEF.

Skinner, B. F. (1948). "Superstition" in the pigeon. *Journal of Experimental Psychology*, *38*, 168–172.

Slavin, R., & Madden, N. (2011). Measures inherent to treatments in program effectiveness reviews. *Journal of Research on Educational Effectiveness*, *4*, 370–380. https://doi.org/10.1080/19345747.2011.558986

Spilt, J., Koot, J., & van Lier, P. (2013). For whom does It work? Subgroup differences in the effects of a school-based universal prevention program. *Prevention Science*, *14*, 479–488. https://doi.org/10.1007/s11121-012-0329-7

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

Tingstrom, D., Sterling-Turner, H., & Wilczynski, S. (2006). The Good Behavior Game: 1969-2002. *Behavior Modification*, *30*, 225–253. https://doi.org/10.1177/0145445503261165

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

van Lier, P., Muthén, B., van der Sar, R., & Crijnen, A. (2004). Preventing disruptive behavior in elementary school children: Impact of a universal classroom-based intervention. *Journal of Consulting and Clinical Psychology*, *72*, 467–478. https://doi.org/10.1037/0022-006X.72.3.467

Wehby, J., Maggin, D., Partin, T., & Robertson, R. (2011). The impact of working alliance, social validity, and teacher burnout on implementation fidelity of the Good Behavior Game. *School Mental Health*, *4*, 22–33. https://doi.org/10.1007/s12310-011-9067-4

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

Mean baseline characteristics at individual and school levels

Demographic	School			Student		
	Overall	GBG	UP	Overall	GBG	UP
Size – number of pupils on roll	306.9	298.2	315.4	-	-	-
Sex – proportion of male students	-	-	-	52.6	50.4	54.9
FSM – proportion of pupils eligible for FSM	26.0	27.6	24.5	24.8	27.4	22.8
EAL – proportion of pupils speaking EAL	22.6	22	23.2	27.3	26.2	31
Ethnic Minority – proportion of ethnic minority pupils	32.9	32.4	33.3	33.5	32.8	34.2
SEND – proportion of pupils with SEND	19.5	20.9	18.2	20.3	23.1	17.6

Table 2

Cumulative risk index development

Significant risk factors for behavioral outcomes

Risk factor (RF)	Disruptive	Concentration	Pro-social
School-level			
High urbanicity		✓	
% EAL students	✓		✓
% students conduct problems	✓	✓	✓
Student-level			
Male gender	✓	✓	✓
Summer-born		✓	
FSM eligible	✓	✓	✓
SEND	✓	✓	✓
Looked-after child	✓	✓	✓

Number of students by risk group and trial allocation

Risk Group	N RF	N students	% sample	N GBG	N UP
Disruptive behavior					
Low-risk	0+1	1638	56	762	876
Medium-risk	2+3	1181	40	644	537
High-risk	4+	119	4	82	37
Concentration problems					
Low-risk	0+1	568	19.9	285	283
Medium-risk	2+3	1780	62.2	837	943
High-risk	4+5	513	17.9	303	210
Pro-social behavior					
Low-risk	0+1	1680	55.3	782	898
Medium-risk	2+3	1228	40.4	674	554
High-risk	4+	129	4.2	88	41

Table 3

Mean (standard error) behavior scores by trial group allocation

	GBG arm		UP arm	
	Pre-test score	Post-test score	Pre-test score	Post-test score
Disruptive behavior	1.66 (.022)	1.74 (.025)	1.59 (.022)	1.64 (.023)
Low-risk	1.46 (0.21)	1.59 (0.03)	1.40 (0.20)	1.46 (0.02)
Medium-risk	1.91 (0.37)	1.89 (0.04)	1.91 (0.42)	1.90 (0.05)
High-risk	2.37 (0.10)	2.01 (0.11)	2.32 (0.19)	2.46 (0.20)
Concentration problems	2.57 (.033)	2.54 (.033)	2.53 (.032)	2.49 (.031)
Low-risk	1.94 (.050)	2.06 (.061)	1.91 (.052)	1.88 (.055)
Medium-risk	2.58 (.037)	2.47 (.043)	2.55 (.037)	2.53 (.039)
High-risk	3.33 (.064)	3.08 (.193)	3.40 (.078)	3.26 (.080)
Pro-social behavior	4.94 (.025)	4.81 (.027)	4.97 (.025)	4.93 (.026)
Low-risk	5.16 (.027)	4.96 (.035)	5.21 (.027)	5.15 (.030)
Medium-risk	4.66 (.036)	4.67 (.042)	4.63 (.044)	4.63 (.046)
High-risk	4.25 (.093)	4.53 (.111)	3.93 (.136)	4.13 (.157)

Table 4

Main effects of the GBG on behavior

	Coefficient β	SE	p value
Disruptive behavior			
β0ij = --1.491(0.120); df = 6			
School-level	0.093	0.018	<.001**
Trial group (if GBG)	0.056	0.077	.235
Proportion FSM	0.001	0.003	.370
Size	0.000	0.000	.500
Student-level	0.522	0.015	<.001**
FSM	0.098	0.036	.004*
SEN	0.132	0.040	<.001**
Gender (male)	0.150	0.031	<.001**
Baseline disruptive behavior scores	0.786	0.021	<.001**
Concentration problems			
β0ij = -1.726(0.130); df = 6			
School-level	0.115	0.021	<.001**
Trial group (if GBG)	0.022	0.084	.400
Proportion FSM	0.152	0.035	<.001**
Size	0.000	0.000	.500
Student-level	0.498	0.014	<.001**
FSM	0.004	0.003	.093
SEN	0.256	0.041	<.001**
Gender (male)	0.232	0.030	<.001**
Baseline concentration	0.506	0.015	<.001**

problems scores			
Pro-social behavior			
β0ij = -1.963(0.193); df = 6			
School-level	0.195	0.035	<.001**
Trial group (if GBG)	-0.108	0.108	.160
Proportion FSM	-0.003	0.004	.160
Size	-0.000	0.000	.500
Student-level	0.604	0.017	<.001**
FSM	-0.136	0.039	<.001**
SEN	-0.278	0.044	<.001**
Gender (male)	-0.171	0.032	<.001**
Baseline pro-social behavior scores	0.476	0.021	<.001**

* p<.05; ** p<.01

FSM = free school meals; SEN = special educational needs

Table 5

Interaction between trial group allocation and program differentiation

	Coefficient β	SE	p value
Disruptive behavior			
β0ij = -1.369(0.087); df = 4			
School-level	0.092	0.018	<.001**
Trial group (if GBG)	0.042	0.109	.351
GBG* high differentiation	0.028	0.154	.428
High differentiation	-0.022	0.107	.419
Student-level	0.534	0.016	<.001**
Baseline disruptive behavior scores	0.832	0.020	<.001**
Concentration problems			
β0ij = -1.455(0.099); df = 4			
School-level	0.127	0.024	<.001**
Trial group (if GBG)	-0.007	0.125	
High differentiation	-0.078	0.123	.478
GBG* high differentiation	0.065	0.177	.357
Student-level	0.524	0.015	<.001**
Baseline concentration problems scores	0.577	0.014	<.001**
Prosocial skills			
β0ij = -2.690(0.149); df = 4			
School-level	0.194	0.035	<.001**
Trial group (if GBG)	0.012	0.153	.469
High differentiation	0.115	0.152	.226
GBG* high differentiation	-0.269	0.217	.109
Student-level	0.626	0.018	<.001**
Baseline prosocial skills scores	0.545	0.020	<.001**

* p<.05; ** p<.01

FSM = free school meals; SEN = special educational needs

Table 6

Subgroup effects of the GBG on behavior for students at-risk

	Coefficient β	SE	p value
Disruptive behavior			
β0ij = -1.487(0.119); df = 11			
School-level	0.090	0.018	<.001**
Trial group (if GBG)	0.049	0.082	.476
Proportion FSM	0.001	0.003	.370
Size	0.000	0.000	.500
Student-level	0.523	0.015	<.001**
FSM	0.118	0.044	.005*
SEN	0.159	0.049	<.001**
Gender (male)	0.170	0.039	<.001**
Baseline disruptive behavior scores	0.785	0.021	<.001**
Risk group:			
Low-risk	◊	◊	◊
Medium-risk	-0.062	0.066	.175
High-risk	-0.005	0.167	.488
GBG*medium-risk	0.055	0.070	.217
GBG*high-risk	-0.234	0.181	.100
Concentration problems			
β0ij = -1.766(0.135); df = 11			
School-level	0.115	0.022	<.001**
Trial group (if GBG)	0.090	0.104	.195

Proportion FSM	0.004	0.003	.093
Size	0.000	0.000	.500
Student-level	0.497	0.014	<.001**
FSM	0.137	0.041	<.001**
SEN	0.237	0.046	<.001**
Gender (male)	0.215	0.037	<.001**
Baseline concentration problems scores	0.507	0.015	<.001**
Risk group:			
Low-risk	◊	◊	◊
Medium-risk	0.068	0.060	.130
High-risk	0.108	0.100	.142
GBG*medium-risk	-0.086	0.078	.137
GBG*high-risk	-0.098	0.106	.179
Prosocial skills			
β0ij = -1.962(0.194); df = 11			
School-level	0.193	0.035	<.001**
Trial group (if GBG)	-0.108	0.112	.169
Proportion FSM	-0.003	0.004	.228
Size	-0.000	0.000	.500
Student-level	0.603	0.017	<.001**
FSM	-0.155	0.048	<.001**
SEN	-0.301	0.054	<.001**
Gender (male)	-0.185	0.042	<.001**
Baseline prosocial	0.476	0.021	<.001**

skills scores			
Risk group:			
Low-risk	◊	◊	◊
Medium-risk	0.034	0.073	.321
High-risk	0.019	0.185	.459
GBG*medium-risk	-0.027	0.076	.362
GBG*high-risk	0.191	0.195	.165

* p<.05; ** p<.01

◊ reference category

FSM = free school meals; SEN = special educational needs

Table 7

Sensitivity analysis for effects of the GBG on behavior for students at-risk (continuous CR score)

	Coefficient β	SE	p value
Disruptive behavior			
β0ij = -1.457(0.120); df = 9			
School-level	0.087	0.017	<.001**
Trial group (if GBG)	0.070	0.090	.230
Proportion FSM	0.001	0.003	.370
Size	0.000	0.000	.500
Student-level	0.522	0.015	<.001**
FSM	0.221	0.066	<.001**
SEN	0.254	0.059	<.001**
Gender (male)	0.277	0.064	<.001**
Baseline disruptive behavior scores	0.789	0.022	<.001**
CR score	-0.129	0.060	.017*
GBG*CR score	0.004	0.034	.453
Concentration problems			
β0ij = -1.757(0.140); df = 9			
School-level	0.111	0.021	<.001**
Trial group (if GBG)	0.142	0.112	.104
Proportion FSM	0.004	0.003	.093
Size	0.000	0.000	.500
Student-level	0.497	0.014	<.001**
FSM	0.172	0.046	<.001**

SEN	0.279	0.051	<.001**
Gender (male)	0.253	0.042	<.001**
Baseline concentration problems scores	0.507	0.015	<.001**
CR score	0.003	0.034	.460
GBG*CR score	-0.048	0.030	.057
Prosocial skills β0ij = -1.951(0.196); df = 9			
School-level	0.189	0.034	<.001**
Trial group (if GBG)	-0.165	0.120	.087
Proportion FSM	-0.003	0.004	.228
Size	-0.000	0.000	.500
Student-level	0.604	0.017	<.001**
FSM	-0.184	0.087	.019*
SEN	-0.329	0.089	<.001**
Gender (male)	-0.220	0.085	<.001**
Baseline prosocial skills scores	0.576	0.022	<.001**
CR score	0.033	0.081	.343
GBG*CR score	0.034	0.038	.187

* p<.05; ** p<.01

FSM = free school meals; SEN = special educational needs

Table 8

Sensitivity analysis for effects of the GBG on behavior for students at-risk (quadratic)

	Coefficient β	SE	p value
Disruptive behavior			
β0ij = -1.541(0.119); df = 9			
School-level	0.089	0.017	<.001**
Trial group (if GBG)	0.162	0.080	.023*
Proportion FSM	0.001	0.003	.370
Size	0.000	0.000	.500
Student-level	0.519	0.015	<.001**
FSM	0.104	0.036	2.889
SEN	0.141	0.042	<.001**
Gender (male)	0.162	0.031	<.001**
Baseline disruptive behavior scores	0.786	0.021	<.001**
Quadratic CR	0.045	0.017	.005*
GBG*Quadratic CR	-0.092	0.023	<.001**
Concentration problems			
β0ij = -1.712(0.132); df = 9			
School-level	0.115	0.022	<.001**
Trial group (if GBG)	0.022	0.088	.402
Proportion FSM	0.004	0.003	.093
Size	0.000	0.000	.500
Student-level	0.500	0.015	<.001**
FSM	0.151	0.036	<.001**
SEN	0.257	0.044	<.001**

Gender (male)	0.232	0.030	<.001**
Baseline concentration problems scores	0.504	0.016	<.001**
Quadratic CR	-0.005	0.017	.385
GBG*Quadratic CR	0.002	0.021	.462
Prosocial skills			
β0ij = -1.963(0.194); df = 9			
School-level	0.194	0.035	<.001**
Trial group (if GBG)	-0.132	0.111	.119
Proportion FSM	-0.003	0.004	.228
Size	-0.000	0.000	.500
Student-level	0.603	0.017	<.001**
FSM	-0.145	0.039	<.001**
SEN	-0.293	0.046	<.001**
Gender (male)	-0.171	0.033	<.001**
Baseline prosocial skills scores	0.476	0.021	<.001**
Quadratic CR	0.004	0.018	.413
GBG*Quadratic CR	0.020	0.025	.213

* p<.05; ** p<.01

FSM = free school meals; SEN = special educational needs

Supplemental Files

Table A.1

Change in behavior scores between risk levels

Risk group	Effect size (Cohen's d)
Disruptive behavior	
0-1	0.35
1-2	0.39
2-3	0.36
3-4+	0.19
Concentration problems	
0-1	0.61
1-2	0.45
2-3	0.42
3-4	0.39
4-5+	0.35
Pro-social behavior	
0-1	0.28
1-2	0.45
2-3	0.35
3-4+	0.25

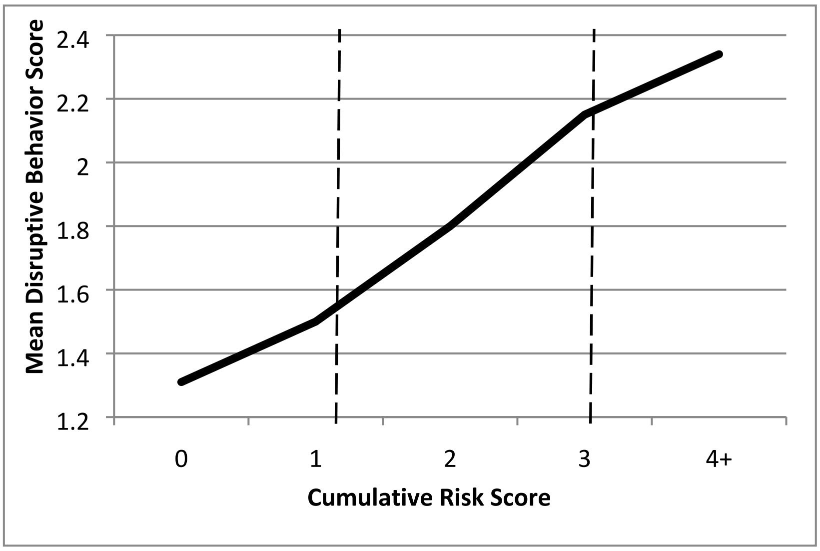

Fig A.1. Line graph demonstrating cumulative risk effect – disruptive behavior

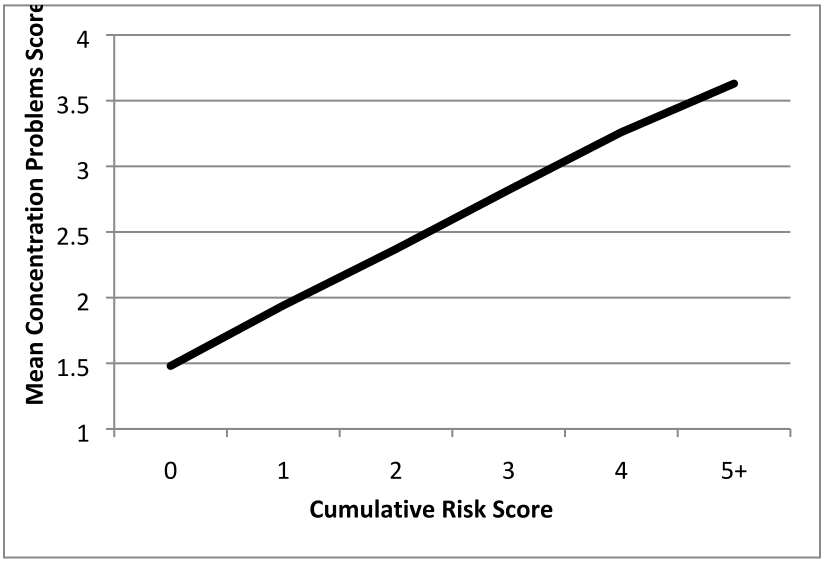

Fig A.2. Line graph demonstrating cumulative risk effect - concentration problems

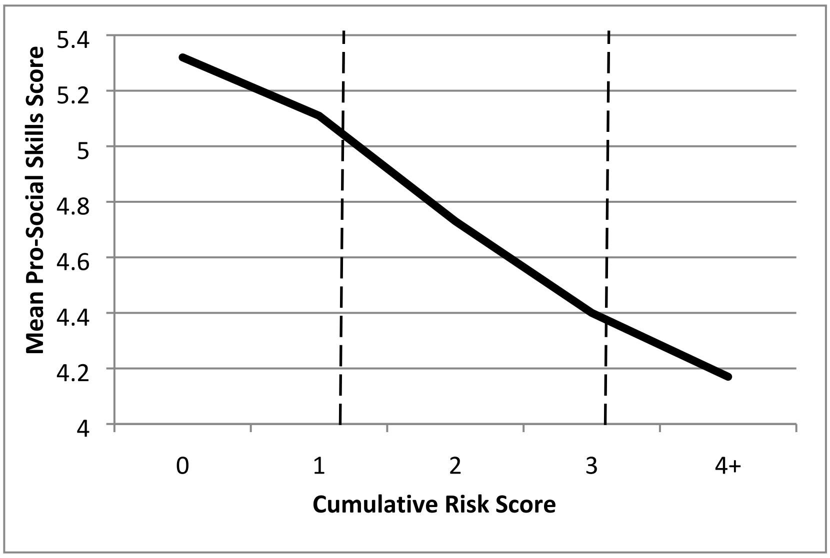

Fig A.3. Line graph demonstrating cumulative risk effect - pro-social behavior

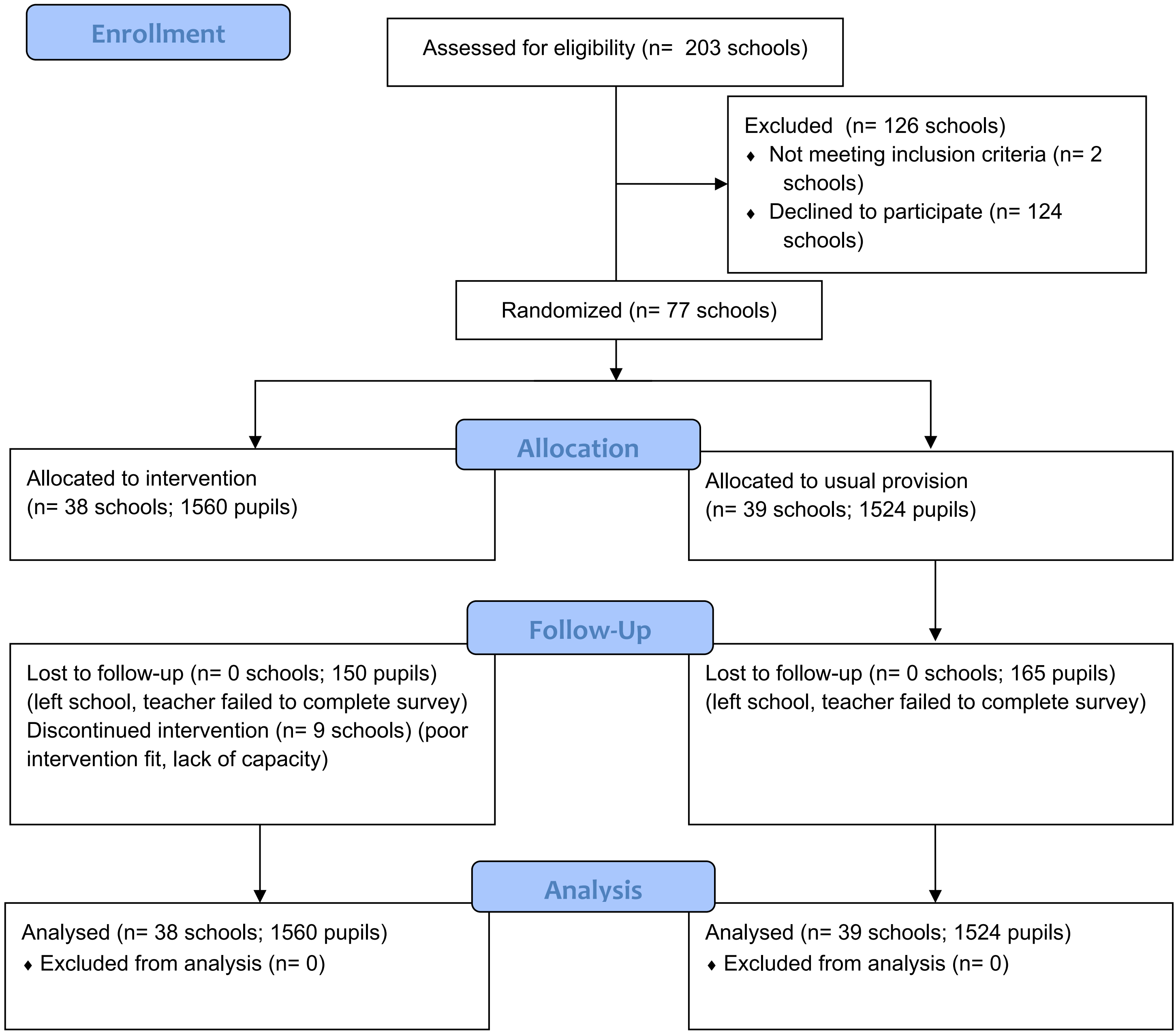

Fig A.4 CONSORT diagram