Humphrey, N, Panayiotou, M, Hennessey, A and Ashworth, E

Treatment effect modifiers in a randomized trial of the good behavior game during middle childhood.

Treatment Effect Modifiers in a Randomized Trial of the Good Behavior Game During Middle Childhood

Neil Humphrey, Margarita Panayiotou, Alexandra Hennessey and Emma Ashworth

Manchester Institute of Education, University of Manchester, UK

Corresponding Author: Professor Neil Humphrey, Manchester Institute of Education, University of Manchester, Oxford Road, Manchester, M13 9PL, United Kingdom. Email: neil.humphrey@manchester.ac.uk

Abstract

Objective: Two key treatment effect modifiers – implementation variability and participant cumulative risk status – are examined as predictors of disruptive behavior outcomes in the context of a large cluster randomized controlled trial of a universal, school-based behavior management intervention. The core components of the Good Behavior Game (GBG) are classroom rules, team membership, monitoring behavior and positive reinforcement. Children work in teams to win the game, which is played alongside a normal classroom activity, during which their teacher monitors infractions to classroom rules. Teams with four or fewer infractions at the end of the game win and are rewarded. **Method:** 77 English primary schools (*N* = 3,084 children, aged 6–7) were randomly assigned to deliver the GBG or continue their usual practice over two years. **Results:** Intent-to-treat analysis found no discernible impact of the intervention on children's disruptive behavior. Additionally, subgroup analyses revealed no differential gains among children at low, moderate or high levels of cumulative risk exposure (CRE). However, complier average causal effect estimation (CACE) using dosage as a compliance marker identified a large, statistically significant intervention effect (d = -1.35) among compliers (>1030 minutes of cumulative intervention exposure). Furthermore, this compliance effect varied by participant CRE, such that children at high and low levels of exposure experienced significantly greater and lesser reductions in disruptive behavior respectively. **Conclusions:** These findings highlight the importance of optimizing implementation and demonstrate the utility of CRE as a theoretically informed approach to subgroup moderator analysis. Implications are discussed and study strengths and limitations are noted.

Keywords: Behavior management, intervention, school, dosage, cumulative risk

Public Health Significance Statements

1. This study provides robust evidence that dosage is a powerful treatment effect modifier in the Good Behavior Game (GBG). To produce meaningful reductions in disruptive behavior, teachers need to play the game for at least 1030 minutes over a two-year period.
2. When playing the GBG, children at different levels of cumulative risk exposure experience differential gains from these higher levels of dosage. Notably, those at the highest levels of risk exposure benefit the most.
3. This study highlights the importance of considering 'how and why' and 'for whom' universal behavior management interventions like the GBG work.

1.

Introduction

By virtue of their wide reach, prolonged period of engagement, and central role in most communities, schools are ideal settings in which to implement universal interventions designed to prevent the development, maintenance or escalation of social, emotional and/or behavioral difficulties among children and young people (Greenberg, 2010). The evidence base is well advanced with respect to the basic question of 'what works' (Tanner-Smith, Durlak, & Marx, 2018). However, our understanding of 'how and why' (e.g., the influence of implementation variability, and change mechanisms underpinning outcomes), 'for whom' (e.g., subgroup moderator effects), 'when' (e.g., timing of intervention effects) and 'at what cost' (e.g., cost-effectiveness) interventions work is considerably less well developed (Durlak, 2015; Farrell, Henry, & Bettencourt, 2013; Greenberg & Abenavoli, 2017). This paper advances knowledge in relation to the moderating effect of implementation variability, participant risk status, and the interaction between them, as predictors of disruptive behavior outcomes in the context of a universal intervention: the Good Behavior Game (GBG) (Ford, Keegan, Poduska, Kellam, & Littman, 2014).

The GBG is an, "interdependent group-oriented contingency management procedure" (Tingstrom, Sterling-Turner, & Wilczynski, 2006, p.225), whose core components are classroom rules, team membership, monitoring behavior and positive reinforcement. Children work in teams to win the GBG in order to access agreed rewards. The game is played alongside a normal classroom activity for a set period of time, during which the class teacher monitors infractions to four rules: (1) we will work quietly[1]; (2) we will be polite to others; (3) we will get out of our seats with permission; and (4) we will follow directions. Teams with four or fewer infractions at the end of the game win and are rewarded

[1] Working quietly is defined by a noise level set in advance by the teacher that is appropriate to the activity in question.

(Donaldson & Wiskow, 2017). Over time, the game evolves in terms of the frequency and duration of play, and the nature and timing of rewards. The GBG is underpinned by behaviorism (e.g., contingency management and the reproduction of rewarded behavior; Skinner, 1945), social learning theory (e.g., learning of appropriate behavior modelled effectively by other team members; Bandura, 1986), and life course/social field theory (LCSFT; e.g., promotion of adaptive processes to enable children to meet social task demands in the classroom; Kellam et al., 2011).

Multiple randomized controlled trials (RCTs) have provided evidence of the positive impact of the GBG on behavior and related outcomes (see Smith et al., 2019, for a recent meta-analysis). It appears to be particularly effective in reducing *disruptive* behavior (e.g., that which disrupts or interrupts activities of others in the classroom such as talking out, getting out of seat, touching others, being disobedient or aggressive). Following a successful pilot (Chan, Foxcroft, Smurthwaite, Coombes, & Allen, 2012), the first RCT of the GBG in England was conducted (Authors, 2018), from which we derive the findings reported herein.

Beyond Intent-to-Treat in School-Based Intervention Research

While it remains the cornerstone of analysis in RCTs, the intent-to-treat (ITT) principle - in which participant data is analyzed uniformly as per randomization, irrespective of whether a given intervention was subsequently received - is increasingly recognized as problematic, particularly in the context of school-based intervention research (Greenberg & Abenavoli, 2017; Peugh, Strotman, McGrady, Rausch, & Kashikar-Zuck, 2017). ITT analysis assumes complete compliance among those who are randomized to receive the intervention, yet decades of research have shown this to be a fantasy; implementation variability is inevitable (Durlak, 2015). Similarly, ITT analysis underappreciates the natural heterogeneity in universal populations with respect to responsiveness to intervention – in other words, some children and young people will experience differential gains following intervention exposure (Greenberg & Abenavoli, 2017). Thus, failing to account for implementation variability and/or individual differences can lead to biased estimates that may underrate the true potential of preventive interventions.

However, traditional approaches through which implementation variability can be accounted for (e.g., "as treated" and "per protocol") are also problematic because they introduce a different source of bias by stripping out data from so-called 'non-compliers' (Sedgwick, 2015). Complier average causal effect estimation (CACE) and related instrumental variable approaches overcome this problem by using data from compliers *and* non-compliers across the intervention *and* control arms of a trial, and means that an unbiased intervention effect estimate that accounts for implementation variability is possible (Peugh et al., 2017). Although this analytical method has been largely ignored in school-based research until very recently (Peugh & Toland, 2017), its application can have important ramifications for the interpretation of intervention effects. For example, in a trial of 'PATHS to Pax', in which the Promoting Alternative Thinking Strategies curriculum and the GBG are delivered in tandem, Bradshaw, Shukla, Pas, Berg, and Ialongo (2020) found that both the presence and magnitude of intervention effects for at-risk students varied between ITT and CACE models. Thus, an ITT intervention effect on social competence grew in size from 0.01 to 0.28, and previously unidentified effects on academic engagement and emotion regulation emerged in CACE models that took account of variability in intervention dosage.

Analysis of subgroup moderator effects presents similar issues with respect to bias. Central to this is the problem of how to robustly investigate individual differences in responsiveness to intervention while avoiding 'data dredging' (Keller, 2019). It is therefore recommended that subgroup analyses are specified in advance, informed by theory and/or research, and include clear specification of the expected direction of effects and population subgroup(s) of interest, using characteristics measured pre-randomization, such as demographic characteristics, individual differences at baseline and/or family factors (Farrell et al., 2013).

To date, the above issues have largely been explored in isolation; that is, researchers have either focused on implementation *or* subgroup moderator effects. Notable exceptions include Aber, Jones, Brown, Chaudry, and Samples (1998) and Ialongo et al. (1999). These studies provide tentative empirical evidence of an interaction between levels of implementation and subgroup characteristics in predicting intervention effects. In other words, how a given intervention is delivered may matter more for particular groups of children. However, how and why we might expect to see such an interaction has not yet been properly articulated – an issue to which we now turn.

Theorizing the Interaction Between Levels of Implementation and Subgroup Characteristics as a Moderator of School-Based Preventive Intervention Effects

Consistent with social-ecological approaches to understanding implementation (e.g., Domitrovich, 2008; Durlak & DuPre, 2008), we argue that the mechanisms through which implementation variability and subgroup characteristics might interact to modify intervention effects are likely to vary by intervention, outcome(s), dimension(s) of implementation, and the salient features of specific subgroup(s). Given this, a 'universal theory' is implausible. Instead, we offer a specific case example focusing on the GBG. Contrasted with a foundational ITT analysis, three hypotheses are proposed. First, we anticipate increased intervention effects on disruptive behavior in the context of higher GBG dosage (H1). Second, we predict intervention effects to vary by participants' risk status (H2), with those at higher levels of cumulative risk exposure (CRE) accruing significantly greater benefit. Third, we expect the magnitude of CRE subgroup intervention effects to vary by dosage (H3); specifically, we envisage that the differential intervention effects predicted in H2 are amplified in the context of higher levels of GBG dosage.

We focus upon children's disruptive behavior because it is a key proximal outcome of the GBG (Chan et al., 2012) and is developmentally significant, being predictive of adult anti-social behavior and related outcomes (e.g., arrest for a violent offence; Hubbard et al., 2006). Our choice of implementation dosage in H1 aligns with the LCSFT perspective underpinning the GBG, in which the process of playing the game socializes the child into the role of the student by explicitly alerting them to (and rewarding them for meeting) important social task demands in the classroom (e.g., paying attention, following directions) at a key transitional stage in their education[2]. This social adaptation process is cumulative in nature; repeated exposure therefore offers increased opportunities for reinforcement, consolidation, and generalization of learned behaviors. Furthermore, dosage is in keeping with the primary motivation for the CACE parameter, which is to determine treatment effects following *receipt* of an intervention (as opposed to the *offer* of an intervention, as in ITT estimation). Finally, other aspects of implementation (e.g., procedural fidelity, >70%; reach, >95%; participant responsiveness, >70%), assessed via independent observation as part of our trial, were routinely high and less variable than dosage (Authors, 2018). Thus, given the requirement for a single indicator in CACE, dosage was selected.

CRE offers a theoretically informed approach to the establishment of subgroup moderator effects in H2. Traditional subgroup analyses examine a single factor in isolation, ignoring the fact that they cluster and co-occur, and meaning that their apparent importance can be over-estimated. The central premise of cumulative risk theory is that the *number* of risk factors to which a child is exposed is a superior predictor of maladaptive outcomes than the *nature* of individual risk factors. This is based on the proposition that the complex and interactional relationships between risk factors produce amplified effects when they

[2] Children aged 6-7 in England are transitioning from Key Stage 1 to Key Stage 2 in primary school; this is marked by a shift in expectations regarding classroom behavior (e.g. increased desk time).

accumulate that disrupt proximal processes of development, leading to dysfunction (Evans, Li, & Whipple, 2013).

However, CRE has been neglected as a marker for subgroup moderator analyses. In the only application of it in the context of a school-based trial to date, the Multisite Violence Prevention Project (2008, 2009) highlighted its utility by demonstrating that effects of the Responding in Peaceful and Positive Ways and Guiding Responsibility and Expectations in Adolescents Today and Tomorrow interventions on middle school students' aggressive behavior varied by their level of CRE. Our prediction of amplified effects at higher levels of CRE (H2) is based on this empirical precedent and extant perspectives on heterogeneity of effects in preventive interventions (Farrell et al., 2013; Greenberg, 2010; Greenberg & Abenavoli, 2017), in particular the 'compensatory effects' hypothesis (McClelland, Tominey, Schmitt, & Duncan, 2017). More specifically, we theorize that the increased behavioral socialization opportunities associated with GBG intervention processes will offset the significant disruption of developmental processes brought about by CRE. Finally, the prediction of multiplicative effects (H3) is based on the notion that the social adaptation process through which the GBG impacts upon behavior is *cumulative* in nature, and those at higher levels of CRE are likely to benefit more from the increased opportunities for reinforcement, consolidation and generalization of learning associated with increased levels of exposure, as this will mitigate against the lack of adaptive socialization in other developmental contexts.

Method

Design

A cluster-RCT design was used (protocol available here: [masked for review]), with 77 participating schools acting as the unit of randomization. The allocation procedure was conducted by an independent trials unit. Adaptive stratification was used to ensure balance across trial arms in the proportion of children eligible for free school meals (FSM) and school size. 38 schools were randomly allocated to the intervention arm, and implemented the GBG (with technical support and assistance) for two years. 39 schools were randomly allocated to the control arm, and continued their usual practice (UP) throughout this period.

Ethical approval was granted by the authors' host institution (Ref: 15126). All schools signed a Memorandum of Agreement confirming their willingness to participate. Consent was sought from parents/carers, of whom 68 (2.2%) exercised their right to opt their children out of the trial. Finally, children were provided with information about the study (including their guarantee of anonymity and right to withdraw) and were asked to give their assent to participate; none declined assent or exercised their right to withdraw from the study.

Participants

Schools

The composition of the trial schools mirrored that of primary schools in England in respect of size and the proportion of children speaking English as an additional language (EAL), but contained significantly larger proportions of children with special educational needs (SEN) and eligible for FSM, in addition to lower rates of absence and attainment (Authors, 2018). GBG and UP schools did not differ significantly with respect to any of these characteristics (Table 1; Authors, 2018).

[Table 1 near here]

Children

The target cohort was children aged 6–7 in participating schools (N = 3,084). Those attending GBG and UP schools did not differ significantly with respect to sex, FSM, EAL, or SEN (Table 1; Authors, 2018).

Measures

Disruptive Behavior

The nine-item disruptive behavior subscale of the Teacher Observation of Children's Adaptation checklist (TOCA-C; Koth, Bradshaw, & Leaf, 2009) requires teachers to read statements reflecting disobedient, disruptive and aggressive behaviors (e.g., "gets angry when provoked") and endorse them on a six-point scale (from *Never* to *Almost Always*) in relation to a given child (item average score range 1-6; higher scores indicate higher frequency of disruptive behaviors). The TOCA-C is internally consistent (all subscales $\alpha > .86$) and has a factor structure that is invariant across sex, race and age (Koth, Bradshaw, & Leaf, 2009). Internal consistency of the subscale in this trial was excellent ($\alpha = .94$ at baseline).

Cumulative Risk Exposure

To calculate CRE, 16 child-level (being male*, young relative age [e.g. summer born], looked-after [e.g. in the care of their Local Authority]*, identified as having a special educational need [SEN]*, eligible for free school meals [FSM]*, minority ethnic group, speaking English as an additional language [EAL], living in a deprived neighbourhood) and school-level (low average academic achievement, high % of children with SEN, high % of EAL children*, low % average attendance, high % child behavior problems*, large school size, urban location, and high % children eligible for FSM) candidate risk variables spanning multiple ecological domains were regressed onto baseline disruptive behavior scores in a hierarchical linear model (Authors, 2020a). Candidate risk factor selection was based on availability and theoretical and/or empirical precedent (for a detailed review, see Authors, 2018). So, for example, being male and/or identified as having SEN at the child-level, and a higher percentage of children eligible for FSM at the school-level, have each been shown to predict behavioral problems (NHS Digital, 2018; Sellström & Bremberg, 2006).

Both fixed (e.g. male) and variable (e.g. identified as having a SEN) factors were included (Furber, Leach, Guy, & Segal, 2017). This approach is consistent with both cumulative risk theory and the compensatory effects hypothesis underpinning our subgroup analysis, wherein the intervention is not theorized as directly ameliorating risk factors themselves, but rather offsetting the significant disruption of developmental processes brought about by CRE.

Each risk factor was coded as 0 = absent or 1 = present; continuous variables were coded as 1 if the score fell at or above the 75th percentile (Authors, 2020a). Those that were statistically significant predictors in said model (denoted by '*' in the preceding text) were summed, creating a cumulative risk score for each participant that represented the number of risk factors to which they were exposed (ranging from 0–4+)[3]. The functional form of the relationship between CRE and disruptive behavior scores was then assessed and determined to be nonlinear; of particular note was the evidence of distinct elbow points (indicative of 'threshold' effects) between 1 and 2, and 3 and 4+ risk factors. Accordingly, for the subgroup moderator analyses reported herein, participants exposed to 0 or 1 risk factors (*n* = 1,680, 54.5%) were classified as low CRE, those exposed to 2 or 3 (*n* = 1,228, 39.8%) as moderate CRE, and 4+ (*n* = 129, 4.2%) as high CRE. Risk data were missing for the remaining 47 (1.5%) participants.

Implementation

An online scoreboard was developed as part of the trial that automatically recorded the duration and frequency of game play, and allowed teachers to note infractions. This minimized data burden, improved accuracy and guarded against the bias associated with self-reported implementation data (Elswick, Casey, Zanskas, Black, & Schnell, 2016). Data generated were used to ascertain cumulative intervention intensity (Warren, Fey, & Yoder, 2007), with dosage treated as a continuous variable representing total number of minutes' exposure across the two years. As noted earlier, this approach to defining compliance is

[3] As the number of risks increased, the proportion of participants decreased; thus, consistent with established practice in cumulative risk research, children exposed to 4, 5 or 6 risk factors were collapsed into a '4+' category (Authors, 2020a).

justified given that the primary motivation for the CACE parameter is to determine treatment effects following *receipt* of an intervention, and that the social adaptation process of the GBG is theorized to be *cumulative* in nature. Other candidate compliance variables do not provide this information. For example, fidelity data may provide insights into the extent to which a teacher has adhered to prescribed intervention procedures, but tells us nothing about the frequency with which these procedures have been implemented. These are distinct dimensions of implementation, and indeed were weakly correlated (≈.29) in the current study.

The distribution of total minutes of implementation did not deviate substantially from normality (e.g., skew = 1.07, kurtosis = 1.54; both values comfortably below the respective thresholds of 2 and 7 that would indicate substantial deviation; Kim, 2013). The GBG was implemented twice per week on average in the first year of the trial, but this reduced somewhat in the second year; average game duration in both years was approximately 15 minutes (Table 2). Additionally, nine GBG schools formally ceased implementation prior to the conclusion of the trial (though their dosage data are included in the above estimates). Overall, dosage was lower than that reported in some other GBG trials (e.g., Bradshaw et al., 2020). However, these previous trials have relied on teachers' self-reported implementation data, which is known to exhibit substantial positive bias, meaning it likely overestimates actual levels of implementation (Hansen, Pankratz, & Bishop, 2014). Furthermore, as noted by Becker, Bradshaw, Domitrovich, and Ialongo (2013), there is no empirically established benchmark for what constitutes a 'minimally effective dose' of the GBG.

Covariates and Compliance Predictors

Several school-level (e.g., school size, proportion of children eligible for FSM, proportion of children speaking EAL), and child-level (e.g., sex, FSM eligibility, SEN status, concentration problems, pro-social behavior) variables were used as covariates and compliance predictors in the ITT and CACE analyses. These variables were included in order to increase statistical power to detect intervention effects, align with the 'analyze as you randomize' principle (in the case of school size and proportion of children eligible for FSM), account for the influence of known correlates of disruptive behavior, and produce more robust compliance classes and CACE estimates. Although some covariates were also used to construct the CRE score noted above, none were correlated with it above .54; hence, collinearity was not a concern in the subgroup moderator analyses. Furthermore, the inclusion of these covariates created consistency between the ITT and subgroup moderator analyses, the latter being an extension of the former, thereby facilitating direct comparison between the two models.

School-level data were taken from the Department for Education performance table data and child-level data were extracted from the National Pupil Database (NPD), with the exception of concentration problems and pro-social behavior, which were derived from the TOCA-C at baseline.

Analysis

Intent to Treat and Subgroup Moderator Analyses

Multilevel models with fixed slopes and random intercepts were fitted in Mplus 8.3 in view of the hierarchical and clustered nature of the dataset. Fixed slopes were used because there was no evidence that would lead us to expect our baseline to have different predictive relationships with the outcome for each cluster/school (as in a random slopes model). Child was treated as Level 1 and schools as Level 2. Classroom was not treated as a level in our analyses, as information on class membership (i.e., who belonged to which class) was not available for the *control* schools. This is because the main study analyses did not require this information (that is, the ITT analysis involved determination of the effect of a school level variable (GBG vs control) on child level outcomes), and we were conscious of the data burden on schools in the control arm. ITT models included school size, % FSM, % EAL, and trial group as explanatory variables at the school level. Sex, FSM eligibility, SEN status, and baseline concentration problems, pro-social behavior, and disruptive behavior were fitted at the child level, with two-year follow-up disruptive behavior problems as the response variable.

Subgroup moderator analyses extended the ITT models to include cumulative risk exposure at the child level and cross-level interaction terms (e.g., trial group*CRE level). These interaction terms were considered alongside the direct effects of the explanatory variables (Hox, Moerbeek, & de Schoot, 2018) and were interpreted as demonstrating the extent of differential gain among those in the subgroup (e.g., high CRE) in the intervention (compared to usual practice) compared to those not in the subgroup (e.g., low/moderate CRE) (Hancock, Kjaer, Korsholm, & Kent, 2013). More specifically, the beta coefficient was interpreted as the effect modifier size. An interaction of 2 points would indicate, for instance, that those in a given risk subgroup receiving the intervention would benefit by 2 more or less points than those not in said subgroup (Hancock et al., 2013). Given the expected negative relationship between the intervention and disruptive behavior, a *positive* interaction effect in our case would indicate GBG to be less beneficial for those in the given risk subgroup, while a *negative* effect would suggest greater benefits. Three additional models were fitted, one for each subgroup of CRE (low, moderate, high), using a binary variable where 1 corresponded to the focal subgroup (e.g., high CRE) and 0 to the remaining two subgroups (e.g., low/moderate CRE). This was an important modeling decision, particularly for the moderate CRE group (vs. low/high), as it allowed us to examine the tenability of a so-called 'Goldilocks' effect. In other words, the GBG might not be *necessary* for those at low levels of CRE and may be *insufficient* for those at high levels of CRE, but could feasibly trigger behavioral change among those at moderate levels of CRE (Muthén et al., 2002).

CACE Assumptions and Analysis

All CACE analyses were undertaken in MPlus 8.3, the syntax for which can be found in the supplementary materials accompanying the paper. Given that compliance information is missing for the control group, it is treated as a latent (unknown) variable and CACE is estimated probabilistically through mixture modeling, using robust maximum likelihood (MLR) estimation and expectation maximization algorithm, which enables the estimation of the latent variable (Muthén & Muthén, 2017). In other words, individuals in the control schools are classified as compliers or non-compliers, had they been randomized to receive the intervention. This is estimated based on the compliance information that is available for the intervention group and the response distribution information of the sample (Peugh et al., 2017). Following guidance (Jo, Asparouhov, Muthén, Ialongo, & Brown, 2008; Panayiotou et al., 2019), CACE analysis was conducted as multilevel mixture modeling with high starting values (4000 1000) to ensure that the best loglikelihood was achieved. As with the ITT models, school was treated as the unit of randomization (Level 2) and CACE was therefore conducted at the school level.

For the estimation of CACE models we were confident that 1) assignment to the intervention groups was random (Holland, 1988); 2) the assumption of the stable unit treatment value (SUTVA) was met due to the cluster level randomization (i.e., there was no contamination); and, 3) there were no “defiers” or “always-takers”, as the control schools did not have access to GBG (see Angrist, Imbens, and Rubin [1996] for causal inference with CACE). Given the arbitrary thresholds used to define compliance to the intervention (below), we were, however, less confident about the exclusion restriction, which assumes that the intervention effect is zero for non-compliers. Indeed, GBG could still be effective for children in classrooms where it is played less. Although the inclusion of strong predictors can reduce the impact of the exclusion restriction violation, sensitivity analyses were conducted (assuming additivity of treatment effects), where this assumption was relaxed and intervention effects for non-compliers (NACE) were estimated in order to assess the tenability of this assumption (see Model B in Jo, 2002).

Compliance. While the minutes played were recorded at the teacher/classroom level, we were unable to model this as a higher level in our models, as information on the class membership for the control schools was not available. This meant that dosage data needed to be aggregated to the school-level or disaggregated to the child-level. Consistent with our previous research (Authors, 2019) and following expert consultation (Booil Jo and Linda Muthen, personal correspondence, August 2018) we opted for the latter, for three reasons. First, given the limited work done within multilevel CACE, we wanted to follow as much as possible the simulation by Jo, Asparouhov, Muthén, Ialongo, and Brown (2008), which treated implementation as a Level 1 variable. Second, the efficiency of CACE models in which compliance is a Level 2 variable is unclear, and aggregating to the school-level would lead to loss of information (Hox et al., 2018). Third, it was theoretically consistent to treat dosage as a child-level variable given that even though it was decided and recorded by the teachers (e.g., using the online scoreboard) it represented the level of dosage to which children had access. This is typical in educational research where, as Jo, Asparouhov, Muthén, Ialongo, and Brown (2008) suggest, participants, “do not have much room for independent decision on compliance” (p.17).

Compliance was therefore disaggregated to the child-level and was allowed to vary in both levels. For the identification of the latent compliance variable, it was necessary to dichotomize the dosage variable into compliers (score of 1) and non-compliers (score of 0). Given the absence of an established dosage threshold for GBG (Becker et al, 2013), we conducted sensitivity analyses following other studies (Berg et al., 2017) in which compliance was defined in two ways: 1) classrooms that fell above the 50th percentile (1030 minutes) were deemed to be moderate compliers (n_{child} = 672, 43.1%); 2) classrooms that fell above the 75th percentile (1348 minutes) were considered high compliers (n_{child} = 333, 21.3%).

Subgroup moderator analyses. As with ITT, CACE models were extended to include subgroup moderator effects. While interaction terms are commonly used in multilevel modeling for the identification of treatment subgroup effects, this has received no empirical support in multilevel mixture modeling, although recent evidence supports its use in single-level CACE (Nagengast et al., 2018). This stage of analysis was therefore exploratory in nature and results are taken to be indicative rather than conclusive. To accommodate interaction effects in multilevel CACE, several issues were considered. First, given that random slopes are not possible in a multilevel mixture framework, interaction effects were created through multiplication using the DEFINE option in Mplus (Trial group*CRE) and were modelled as child-level predictors. Second, following the exclusion restriction assumption, the main effects but also the interaction effects were set to zero in non-compliers (see Model A in Jo, 2002). However, given that this assumption was less likely to hold, the exclusion restriction was relaxed to also examine its tenability (per Model C in Jo, 2002). Third, given the reduced power observed in studies with interaction effects (Brookes et al., 2004), this analysis was considered only for the moderate compliance model, where the sample size was larger. Finally, given that multilevel CACE models are computationally heavy, the binary CRE variable was centered to the cluster mean, as this is recommended for cross-level interactions (Enders & Tofighi, 2007), while it can also aid with the computation of complicated models (Hayes, 2005). Indeed, preliminary evidence from CACE models without centering indicated substantially inflated standard errors. For consistency, cluster-centering was also applied to the ITT subgroup models.

Effect Size Calculation and Interpretation

An effect size comparable to Cohen's *d* (Cohen, 1992) was calculated in instances where a statistically significant intervention effect was observed using the formula $d = b/\sigma_T$, where b represents the unstandardized treatment beta effect and σ_T indicates the total standard deviation of the outcome variable ($\sigma_{school} + \sigma_{child}$) (Hedges, 2007). For the CACE models specifically, σ_T corresponded to that of the complier class. The empirical distribution of universal school-based prevention program effects (Tanner-Smith et al., 2018), alongside meta-analytic evidence of the average effects of behavior management strategies more specifically (including the GBG; Korpershoek, Harms, de Boer, van Kuijk, & Doolaard, 2016), was used to guide our interpretation.

Results

18.5% of children had data missing at follow-up, in cases where they had left the school (12.6%) or teachers had failed to provide post-test behavior data (5.9%) (see CONSORT diagram in Figure 1). Missingness (yes/no) was used as the response variable in a logistic regression, with other study data as explanatory variables (e.g., sex, FSM eligibility, SEN, TOCA scores at baseline, and at-risk of conduct problems at baseline). SEN status (β = 0.310, p <.05) and baseline pro-social behavior score (β = -0.282, p <.01) both predicted missingness. Accordingly, MLR with full information (FIML) was used for the ITT (including subgroup moderator extension – Table 3) and main CACE models (Table 4) under the assumption of data missing at random. Using FIML for the subgroup moderator extension of CACE models (Table 5) and the NACE models (supplementary materials) would, however, have been computationally expensive, as these required up to seven dimensions of integration, which is more than the recommended maximum of five (Muthen & Muthen, 1998–2017). We therefore used listwise deletion for these models, which we acknowledge as a limitation of the study (see Discussion).

[Figure 1 near here]

Intent to Treat and Subgroup Moderator Models

The main ITT analysis, controlling for child-level and school-level covariates (Table 3), revealed no discernible effect of the GBG on children's disruptive behavior ($\beta = .22$, $p > .05$). Extension of the ITT model to include cross-level interaction terms for subgroup moderator analyses demonstrated no significant differential gains among those at low ($\beta = -.01$, $p > .05$), moderate ($\beta = .03$, $p > .05$), or high ($\beta = -.05$, $p > .05$) levels of CRE.

[Table 3 near here]

CACE, NACE and Subgroup Moderator Models

Moderate and high compliance CACE models are reported in Table 4 and moderate compliance CACE subgroup analyses are reported in Table 5. The former estimate intervention effects accounting for (moderate or high) dosage, while the latter is an extension of the moderate CACE model, in which subgroup moderator effects are examined for children at low, moderate and high levels of CRE. All models had high entropy values and posterior probabilities, while none of the classes had less than 1% of total count, indicating an acceptable solution (Jung & Wickrama, 2008). Intra-cluster correlation coefficients (ICCs) were as follows for the outcome: ICC_{YC} (compliers) = .04, ICC_{YN} (non-compliers) = .13; and for compliance: ICC_C = .97 (moderate) and .99 (high). Complier and non-complier means were .5 standard deviations apart. Drawing on Jo et al. (2008; specifically, Figure 3B), we can therefore conclude that variance misestimation would be low in the current study and coverage would be at acceptable levels (around .8), minimizing the likelihood of biased estimates.

After accounting for child-level and school-level covariates, a large, statistically significant CACE intervention effect was identified in the moderate compliance model ($\beta = -1.72$, $p < .001$, d = -1.35). This effect remained relatively stable in magnitude in the high compliance model ($\beta = -1.75$, $p < .05$, d = -1.14), indicating no additional benefits of increased dosage beyond those accrued through moderate compliance. Upon relaxing the exclusion restriction criterion, CACE effects remained large ($d_{moderate}$ = -1.25; d_{high} = -0.99); however, small positive NACE effects were observed for non-compliers in both moderate (β = .85, p < .01, d = 0.38) and high (β = .80, p < .01, d = 0.31) compliance models, indicating iatrogenic effects for those that did not comply. For NACE sensitivity analyses, see supplementary Table S1.

Extension of the moderate compliance model to include cross-level interaction terms for subgroup moderator analyses demonstrated a significant positive interaction between trial group and low CRE (β = .41, b = .83, p <.001); the corresponding main effect remained large in this extended model (β = -1.84, p <.001, d = -1.77), and the individual effect of risk was significant and negative. Conversely, a significant negative interaction was identified for the high CRE group (β = -.24, b = -1.21, p <.01), with stable main trial effects (β = -1.75, p <.001, d = -1.17), and a positive risk effect (β = .81, b = .81, p <.05). No significant interaction effects were identified for the moderate CRE group.

A similar pattern to that above was observed following the relaxation of the exclusion restriction assumption in the moderate compliance model (see Table S2 in supplementary material): CACE effects remained large and negative, while positive NACE effects were observed for non-compliers in all three risk groups (albeit non-significant for the moderate CRE group). Interaction effects were significant for compliers only and similar to the previous findings (Table 5), positive (β = .41, b = .78, p < .001) and negative (β = -.22, b = -1.14, p < .001) interaction effects were observed for the low and high CRE groups, respectively. Unlike previous analyses, however, when the exclusion restriction was relaxed, a significant negative interaction was identified between trial group and moderate CRE (β = -.27, b = -.56, p < .01; main effect β = -1.46, p < .001, d = -.93); the direct effect from risk to outcome was significant and positive.

[Tables 4 and 5 near here]

Predictors of compliance

No school-level characteristics predicted compliance. For the child-level covariates, in the main moderate compliance model, teachers were less likely to comply in classes with a higher percentage of children with SEN (b = -.64, $p < .05$, OR = .53, $p < .01$). For the high compliance model, teachers were more likely to comply in classrooms with lower levels of concentration problems (b = -.30, $p = .06$, OR = .75, $p < .05$). Finally, both higher percentage of SEN and disruptive behavior problem scores were significant predictors of reduced compliance in the low (SEN b = -.62, OR =.54; Disruptive b = -.54, OR = .58) and moderate risk (SEN b = -.66, OR =.52; Disruptive b = -.53, OR = .59) moderate compliance models, whereas for high risk CACE, only SEN was a significant predictor (b = -.61, OR = .55).

Discussion

The aim of the current study was to examine the moderating influence of implementation variability (dosage), participant characteristics (CRE), and the interaction between them, as predictors of disruptive behavior outcomes in the context of a large randomized trial of the GBG. Drawing upon extant theory and research, we predicted increased intervention effects in the context of higher GBG dosage (H1). Differential gains among children at varying levels of CRE were also anticipated (H2). Finally, we hypothesized larger effects to be generated through the interaction between dosage and CRE levels (H3). H1 was fully supported - null results in our ITT model contrasted sharply with large, statistically significant intervention effects in the moderate and high compliance CACE models. Contrary to our H2 predictions, we found no evidence of differential gains among participants at different levels of CRE when the ITT model was extended to include subgroup moderator analyses. However, H3 was supported; extension of our CACE models to include subgroup moderator analyses revealed that children at high and low CRE levels experienced significantly greater and lesser reductions in disruptive behavior respectively. Sensitivity analyses, where the exclusion restriction assumption was relaxed, further supported the security of our findings, as intervention effects remained stable. However, iatrogenic or demoralisation effects (as in Connell, 2009; Jo, 2002) were found for non-compliers, such that those that played the game for less than 1030 minutes over the two-year trial period reported increases in disruptive behavior. These findings and those reported elsewhere (Connell, 2009; Jo, 2002) also highlight the challenges associated with the CACE assumptions, as the intervention effects were not zero for non-compliers, as the exclusion restriction assumes. The tenability of this assumption should, therefore, be tested where possible, following appropriate estimation techniques (see Jo, 2002).

The stark contrast between our ITT and CACE findings (H1) underscores the importance of using robust methods to account for implementation variability when estimating the effects of school-based interventions (Peugh et al, 2017). This contrast is perhaps best exemplified by the fact that when implemented with sufficient intensity, the GBG can lead to reductions in disruptive behavior of a magnitude that greatly exceeds those produced by other behavior management strategies or universal school-based interventions more generally (Korpershoek et al., 2016; Tanner-Smith et al., 2018). However, when considering only main (ITT) effects, it would certainly not be recommended.

Our CACE models offer the first empirically established benchmark for minimally effective dosage of the GBG (>1030 minutes) in relation to its proximal outcome of disruptive behavior. In addition, the results of our sensitivity analysis (high compliance model, >1348 minutes) demonstrated that further increases in GBG dosage do not lead to great amplification of the magnitude of intervention effects. Taken together, these analyses indicate an optimal range of GBG implementation – between 1030 and 1348 minutes of cumulative intervention exposure over two years – in order to manage behavior most efficiently. This is an issue to which we will return (see 'Implications').

The very similar regression coefficients and large effect sizes in our moderate and high compliance models mirror recent CACE findings for some other school-based interventions (e.g. *PATHS*, Panayiotou, Humphrey & Hennessey, 2019; *Motivation in Mathematics*, Nagengast et al, 2018; *Adolescent Transitions Program*, Connell, 2009). While we are deliberately cautious in drawing meta-inferences given the nascent status of CACE in the study of school-based interventions and the various ways in which 'compliance' may be defined across trials, this emergent pattern of findings does appear to support arguments proposed by Durlak and DuPre (2008) more than a decade ago: not only is the expectation of full implementation unrealistic, it is also unnecessary. This is a point to which we return when discussing the implications of our findings.

Contrary to our initial predictions (H2), we found no evidence of differential gains among children at varying levels of CRE when the ITT models were extended to include subgroup moderator analyses. Given that the only comparable study of a school-based intervention found clear evidence of effects varying by CRE (Multisite Violence Prevention Project, 2008, 2009), what are we to make of this unexpected finding? It could simply be that the change mechanisms through which the GBG impacts disruptive behavior simply do not target those at higher levels of CRE in the manner theorized earlier in this paper. Alternatively, the GBG may work as theorized, but the manner in which CRE was assessed in the current study was somehow flawed or inaccurate, leading to a Type II error (see Strengths and Limitations section below). However, the most likely explanation is perhaps that the GBG works as proposed, and our methodology was sound, but implementation failed to reach sufficient levels to enable the hypothesized effects to be clearly evidenced. An important avenue for future research is therefore to determine whether subgroup moderator effects based on CRE can be established in trials with higher overall levels of implementation.

The pattern of subgroup effects in our moderate compliance model (e.g., significantly greater and lesser reductions in disruptive behavior among participants at higher and lower CRE respectively, compared to the average CRE in their school) was consistent with our predictions (H3), and provides important new evidence that increases our understanding of how treatment effect modifiers may operate in combination to moderate intervention outcomes. Specifically, our findings align with the proposition that the social adaptation process through which the GBG impacts upon behavior is *cumulative* in nature. Thus, those at higher levels of CRE benefit more from the increased opportunities for reinforcement, consolidation and generalization of learning associated with increased levels of exposure, as this mitigates against the lack of adaptive socialization in other developmental contexts. In further support of this, interaction effects were also found in the moderate CRE subgroup, but only when the exclusion assumption was relaxed; those at moderate CRE levels in GBG schools displayed greater decreases in disruptive behavior. These results should, however, be interpreted with caution given their sensitivity to the violation of the exclusion restriction assumption. Sensitivity analysis at high levels of compliance was not performed given the significantly compromised sample size (and consequent reduction in statistical power) in such a model. Thus, future research should seek to establish the extent to the pattern of differential gains by CRE are further intensified at the highest levels of GBG dosage. Such research will require a significantly larger sample than was available in the current study.

The identification of intervention effects varying by CRE in the moderate compliance model adds to the emergent evidence base that demonstrates its utility in subgroup moderator analyses (Multisite Violence Prevention Project, 2008, 2009). When compared to the standard approach of examining differences across one or more socio-demographic variables such as sex or ethnicity (adopted in 54 of 68 studies in Farrell et al's (2013) review of school-based violence prevention studies), CRE offers a more theoretically informed, context-sensitive approach that accounts for the clustering and interaction of risk factors.

Strengths and Limitations

The security of the findings reported in the current study are enhanced by several features. We used a randomized controlled trial design with analyses that took data clustering, implementation variability, and participant risk status into account. The possibility of diffusion/contamination was minimized by the use of cluster randomization, and the random allocation process was undertaken independently of the research team. Trial arms were well balanced at baseline with respect to key observables. Measures of implementation (dosage), participant risk status (CRE) and outcomes (disruptive behavior) were robust and theoretically informed.

However, as noted earlier, although we were able to use full information in our ITT and main CACE models, the subgroup moderator extensions of the CACE models and our NACE models were based on listwise deletion due to the excessive computational demands (higher than the maximum recommended dimensions of integration) of a multilevel FIML CACE model incorporating subgroup moderator effects. Failing to account for missing data can introduce bias and accordingly, said models should therefore be treated with caution. Also, given that classroom membership information for the control schools was not available, we could not model teacher-level characteristics (e.g. self-efficacy of behavior management) as predictors of compliance, and were also unable to explore a 3-level CACE where classroom acted as a cluster level (Child$_{L1}$, Classroom$_{L2}$, School$_{L3}$).

In an ideal scenario, compliance would be measured at the student level, but this was not possible here, as the GBG is a *universal* school-based intervention (e.g., delivered to all children, regardless of need). We maintain that assessing dosage at teacher level is accurate, and great variation at the student level is not expected given the very high levels of reach (>95%) in our study and more broadly by the fact that pupil attendance in English primary schools is uniformly very high (>95%) (HM Government, 2020). Nonetheless, future work could explore ways in which compliance within universal interventions is assessed via both the teacher and their students. From an analytic perspective, this is possible (albeit challenging for large school-based intervention trials) (Schochet & Chiang, 2011). In terms of dosage this might, for example, incorporate daily attendance data into analyses of the kind reported herein (though we note that this may still be flawed since these data measure school attendance on a given day and not whether children were physically present at a particular point in time when a universal intervention was being delivered).

Furthermore, because CACE requires a single indicator, only dosage data were used in our analysis. While dosage was the most appropriate compliance proxy, this did mean that other potentially important implementation dimensions (e.g., procedural fidelity) were neglected. Moreover, our reliance on teacher-reported disruptive behaviour scores via the TOCA-C may have introduced bias, given that trial group allocation was not masked. However, capturing independent (blinded) observational data on over 3,000 children across nearly 80 schools was well beyond the resources available in the trial, and would have created a significant additional burden on the schools themselves. Furthermore, conducting truly blinded observations would be very difficult (if not impossible) given the proliferation of visual artefacts (e.g., GBG classroom rules posters, reward charts and booklets) in intervention classrooms. Finally, although our CRE variable was derived from a wide range of candidate risk variables, data pertaining to other factors such as neonatal complications and familial dysfunction (Evans, Li, & Whipple, 2013) were not available. We therefore recommend that future intervention research involving subgroup moderator analyses based on CRE incorporate a wide-ranging approach to the assessment of risk factors, possibly involving bespoke instruments (as opposed to the secondary analysis undertaken here).

Implications

The optimal range of cumulative intervention intensity revealed in our CACE analyses suggests that modifications to the developer's recommended dosage levels (up to 40 minutes of gameplay, five times per week; Ford et al., 2014) may be necessary. Moderate compliers played the game, on average, 2.2 times per week for approximately 34 minutes, in order to produce the large reductions in disruptive behavior observed in this study. This is well below the number of minutes typically needed for other behavioral interventions, and indicates that the GBG may therefore offer a particularly time-efficient model.

While violation of the exclusion restriction assumption was expected, we found that the impact of the GBG in the context of non-compliance was iatrogenic (e.g. *increases* in disruptive behavior). This finding aligns with that of Owens et al. (2020), who observed reductions in rule violations among students of teachers whose implementation of appropriate behavior management strategies reached or exceed a minimum benchmark following a consultation intervention, but increases among students of teachers whose implementation was inconsistent. Such effects could be the result of a displacement process, wherein existing behavior management approaches were abandoned in favour of the GBG, which was then implemented below a minimally effective dosage. We are cautious, however, in thinking about how literally one might apply these findings, for three reasons. First, replication is obviously required. Second, by focusing on the total amount of intervention exposure, our analysis did not allow us to determine whether frequency or duration of gameplay is most important; this issue should be examined in future research. Third, if teachers were instructed to follow a truncated delivery model, they would likely have to demonstrate full compliance in order to replicate the effects on disruptive behavior observed here.

Although primarily used in order to ensure robust identification of compliers in the control arm of the trial, the establishment of compliance predictors (SEN, disruptive behavior, and concentration problems) also yields practical implications. The proportion of children with SEN was most consistently identified and was always associated with significantly *reduced* likelihood of compliance. One possibility is that, given a multi-tiered system of support, classrooms with higher proportions of children with SEN already benefitted from more intensive Tier 2 behavioral supports (e.g., from teaching assistants), rendering the Tier 1 GBG less necessary and/or in conflict with existing practices from the perspective of participating teachers. This aligns well with a key finding in the qualitative strand of our implementation and process evaluation, whereby teachers reported feeling that the prohibition of interaction with children during gameplay periods was at odds with their inclination to directly support those with SEN to complete the academic activity being undertaken (Authors, 2020b). Thus, some adaptation to the GBG gameplay protocol (e.g., special exception to allow direct support for children with SEN as required during gameplay) may be required in order to optimize implementation for the benefit of all.

The findings of the current study also raise interesting questions in relation to the conceptualization and application of the GBG as a Tier 1 (e.g., universal) strategy. One might, for example, argue that the finding of the greatest benefit being found to those at greatest risk is somewhat contradictory to the conceptual notion of Tier 1 supports. However, as has been noted in the literature (e.g. Farrell, Henry & Bettencourt, 2013; Greenberg & Abenavoli, 2017), universal preventive interventions should not be expected to confer universal benefit. This is particularly the case when one considers our primary outcome of disruptive behavior, as we know that the behavior of the overwhelming majority of children is not a cause for concern (Office for Standards in Education, 2014). Our findings indicate that, when implemented with sufficient levels of dosage, significant benefits are accrued for a subgroup of children – those exposed to higher levels of cumulative risk - who would typically be classed as in need of Tier 2 (e.g., targeted) supports. Given this, the GBG could perhaps be *conceptualized* as a Tier 2 support that is *applied* universally. Thus, even though most of a given class do not ‘need’ the intervention, their participation remains critical in order to for effective socialization behaviors to be modeled for those most at-risk. This view is consistent with the social learning theory underpinnings of the GBG.

Conclusion

This study has demonstrated the importance of intervention compliance, participant CRE, and the interaction between them, as treatment effect modifiers in the Good Behavior Game. In simple terms, we found that higher levels of intervention exposure were critical to the production of reductions in disruptive behavior, but particularly so for those children at high levels of cumulative risk exposure, who accrued significantly greater benefits than their low cumulative risk counterparts in the context of increased compliance. These findings add new, independent and rigorous evidence for the intervention, and by extension, our understanding of how to effectively manage disruptive behavior in the classroom. From a methodological perspective, the study highlights the utility of CACE estimation and CRE as theoretically informed approaches to understanding ‘how and why’ and ‘for whom’ interventions work, and in doing so, demonstrates the value of going beyond ITT.

References

Aber, J. L., Jones, S. M., Brown, J. L., Chaudry, N., & Samples, F. (1998). Resolving conflict creatively: Evaluating the developmental effects of a school-based violence prevention program in neighborhood and classroom context. *Development and Psychopathology*, *10*(2), 187–213. https://doi.org/10.1017/s0954579498001576

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Prentice-Hall.

Becker, K. D., Bradshaw, C. P., Domitrovich, C., & Ialongo, N. S. (2013). Coaching teachers to improve implementation of the good behavior game. *Administration and Policy in Mental Health*, *40*(6), 482–493. https://doi.org/10.1007/s10488-013-0482-8

Bradshaw, C. P., Shukla, K. D., Pas, E. T., Berg, J. K., & Ialongo, N. S. (2020). Using complier average causal effect estimation to examine student outcomes of the PAX Good Behavior Game when integrated with the PATHS Curriculum. *Administration and Policy in Mental Health and Mental Health Services Research*. Online First. https://doi.org/10.1007/s10488-020-01034-1

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

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.apa.org/doiLanding?doi=10.1037%2F0033-2909.112.1.155

Connell, A. M. (2009). Employing complier average causal effect analytic methods to examine effects of randomized encouragement trials. *American Journal of Drug and Alcohol Abuse*, *35*(4), 253–259. https://doi.org/10.1007/s10439-011-0452-9.

Domitrovich, C. (2008). Maximizing the implementation quality of evidence-based preventive interventions in schools: A conceptual framework. *Advances in School Mental Health Promotion*, *1*(3), 6–28. http://www.tandfonline.com/doi/abs/10.1080/1754730X.2008.9715730

Domitrovich, C. E., Gest, S. D., Jones, D., Gill, S., & Sanford Derousie, R. M. (2010). Implementation quality: Lessons learned in the context of the Head Start REDI Trial. *Early Childhood Research Quarterly*, *25*(3), 284–298. https://doi.org/10.1016/j.ecresq.2010.04.001

Donaldson, J. M., & Wiskow, K. M. (2017). The Good Behaviour Game. In B. Teasdale & M. S. Bradley (Eds.), *Preventing Crime and Violence* (pp. 229–241). Springer. https://doi.org/10.1007/978-3-319-44124-5

Durlak, J. A. (2015). Studying program implementation is not easy but it is essential. *Prevention Science*, *16*(8), 1123-1127. https://doi.org/10.1007/s11121-015-0606-3

Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: a review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology*, *41*(3–4), 327–350. https://doi.org/10.1007/s10464-008-9165-0

Elswick, S., Casey, L. B., Zanskas, S., Black, T., & Schnell, R. (2016). Effective data collection modalities utilized in monitoring the good behavior game: Technology-based data collection versus hand collected data. *Computers in Human Behavior*, *54*(1), 158–169. https://doi.org/10.1016/j.chb.2015.07.059

Enders, C., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, *12*(2), 121–138. https://doi.org/10.1037/1082-989X.12.2.121

Evans, G., Li, D., & Whipple, S. (2013). Cumulative risk and child development. *Psychological Bulletin*, *139*(6), 1342–1396. https://doi.org/10.1037/a0031808

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

Ford, C., Keegan, N., Poduska, J., Kellam, S., & Littman, J. (2014). *Good Behaviour Game implementation manual*. American Institutes for Research.

Furber, G., Leach, M., Guy, S., & Segal, L. (2017). Developing a broad categorisation scheme to describe risk factors for mental illness, for use in prevention policy and planning. *Australian and New Zealand Journal of Psychiatry*, *51*(3), 230–240. https://doi.org/10.1177/0004867416642844

Greenberg, M. T. (2010). School-based prevention: current status and future challenges. *Effective Education*, *2*(1), 27–52. https://doi.org/10.1080/19415531003616862

Greenberg, Mark T., & Abenavoli, R. (2017). Universal interventions: Fully exploring their impacts and potential to produce population-level impacts. *Journal of Research on Educational Effectiveness*, *10*(1), 40–67. https://doi.org/10.1080/19345747.2016.1246632

Hancock, M. J., Kjaer, P., Korsholm, L., & Kent, P. (2013). Interpretation of subgroup effects in published trials. *Physical Therapy*, *93*(6), 852–859. https://doi.org/10.2522/ptj.20120296

Hansen, W. B., Pankratz, M. M., & Bishop, D. C. (2014). Differences in observers' and teachers' fidelity assessments. *Journal of Primary Prevention*, *35*(5), 297–308. https://doi.org/10.1007/s10935-014-0351-6

Hayes, A. F. (2005). *Statistical methods for communication science*. Routledge.

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, *32*(4), 341–370. https://doi.org/10.3102/1076998606298043

Her Majesty's Government (2020). *Pupil absence in schools in England: autumn term 2019/20*. Accessed at: https://explore-education-statistics.service.gov.uk/find-statistics/pupil-absence-in-schools-in-england-autumn-term

Hox, J. ., Moerbeek, M., & de Schoot, R. (2018). *Multilevel analysis: Techniques and applications* (3rd ed.). Routledge.

Hubbard, S., Masyn, K. E., Poduska, J., Schaeffer, C. M., Petras, H., Ialongo, N., & Kellam, S. (2006). A comparison of girls' and boys' aggressive-disruptive behavior trajectories across elementary school: Prediction to young adult antisocial outcomes. *Journal of Consulting and Clinical Psychology*, *74*(3), 500–510. https://doi.org/10.1037/0022-006x.74.3.500

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

Ialongo, N. S., Werthamer, L., Kellam, S. G., Brown, C. H., Wang, S., & Lin, Y. (1999). Proximal impact of two first-grade preventive interventions on the early risk behaviors for later substance abuse, depression, and antisocial behavior. *American Journal of Community Psychology*, *27*(5), 599–641. https://doi.org/10.1023/A:1022137920532

Jo, B. (2002). Estimation of intervention effects with noncompliance: Alternative model specifications. *Journal of Educational and Behavioral Statistics*, *27*(4), 385–409. https://www.jstor.org/stable/3648125

Jo, B., Asparouhov, T., Muthén, B. O., Ialongo, N. S., & Brown, C. H. (2008). Cluster randomized trials with treatment noncompliance. *Psychological Methods*, *13*(1), 1–18. https://doi.org/10.1037/1082-989X.13.1.1

Jung, T., & Wickrama, K. A. S. (2008). An introduction to latent class growth analysis and growth mixture modeling. *Social and Personality Psychology Compass*, *2*(1), 302–317. https://doi.org/10.1111/j.1751-9004.2007.00054.x

Kellam, S. G., Mackenzie, A. C. L., Brown, C. H., Poduska, J. M., Wang, W., Petras, H., & Wilcox, H. C. (2011). The Good Behavior Game and the future of prevention and treatment. *Addiction Science & Clinical Practice*, *6*(1), 73–84. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3188824&tool=pmcentrez&rendertype=abstract

Keller, F. (2019). Subgroup analysis: "What works best for whom and why?" In Z. Sloboda, H. Petras, E. Robertson, & R. Hingson (Eds.), *Prevention of substance use* (pp. 247–261). Springer.

Kim, H.-Y. (2013). Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis. *Restorative Dentistry & Endodontics*, *38*(1), 52. https://doi.org/10.5395/rde.2013.38.1.52

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*(3), 1–38. https://doi.org/10.3102/0034654315626799

Koth, C. W., Bradshaw, C. P., & Leaf, P. J. (2009). Teacher Observation of Classroom Adaptation-Checklist: Development and factor Structure. *Measurement and Evaluation in Counseling and Development*, *42*(1), 15–30. https://doi.org/10.1177/0748175609333560

McClelland, M. M., Tominey, S. L., Schmitt, S. A., & Duncan, R. (2017). SEL interventions in early childhood. *The Future of Children*, *27*(1), 33–47. https://files.eric.ed.gov/fulltext/ED590403.pdf

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*(4), 231–244. https://doi.org/10.1371/journal.pone.0178059

Multisite Violence Prevention Project. (2009). The ecological effects of universal and selective violence prevention programs for middle school students: A randomized trial. *Journal of Consulting and Clinical Psychology*, *77*(3), 526–542. https://doi.org/10.1038/mp.2011.182.doi

Muthén, B., Brown, C. H., Masyn, K., Jo, B., Khoo, S.-T., Yang, C.-C., … Liao, J. (2002). General growth mixture modeling for randomized preventive interventions. *Biostatistics*, *3*(4), 459–475. https://doi.org/10.1093/biostatistics/3.4.459

NHS Digital. (2018). *Mental health of children and young people in England, 2017*. NHS Digital.

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

Owens, J. S., Evans, S. W., Coles, E. K., Holdaway, A. S., Himawan, L. K., Mixon, C. S., & Egan, T. E. (2020). Consultation for classroom management and targeted interventions: Examining benchmarks for teacher practices that produce desired change in student behavior. *Journal of Emotional and Behavioral Disorders*, *28*(1), 52–64. https://doi.org/10.1177/1063426618795440

Peugh, J. L., 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

Peugh, J. L., & Toland, M. D. (2017). Psychometric and quantitative methods for school psychology. *Journal of School Psychology*, *60*(2), 5–6. https://doi.org/10.1016/j.jsp.2017.01.001

Schochet, P. Z., & Chiang, H. S. (2011). Estimation and identification of the complier average causal effect parameter in education RCTs. *Journal of Educational and Behavioral Statistics*, *36*(3), 307-345. https://doi.org/10.3102/1076998610375837

Sedgwick, P. (2015). Intention to treat analysis versus per protocol analysis of trial data. *BMJ (Clinical Research Ed.)*, *350*, h681. https://doi.org/10.1136/bmj.h681

Sellström, E., & Bremberg, S. (2006). Is there a “school effect” on pupil outcomes? A review of multilevel studies. *Journal of Epidemiology and Community Health*, *60*(2), 149–155. https://doi.org/10.1136/jech.2005.036707

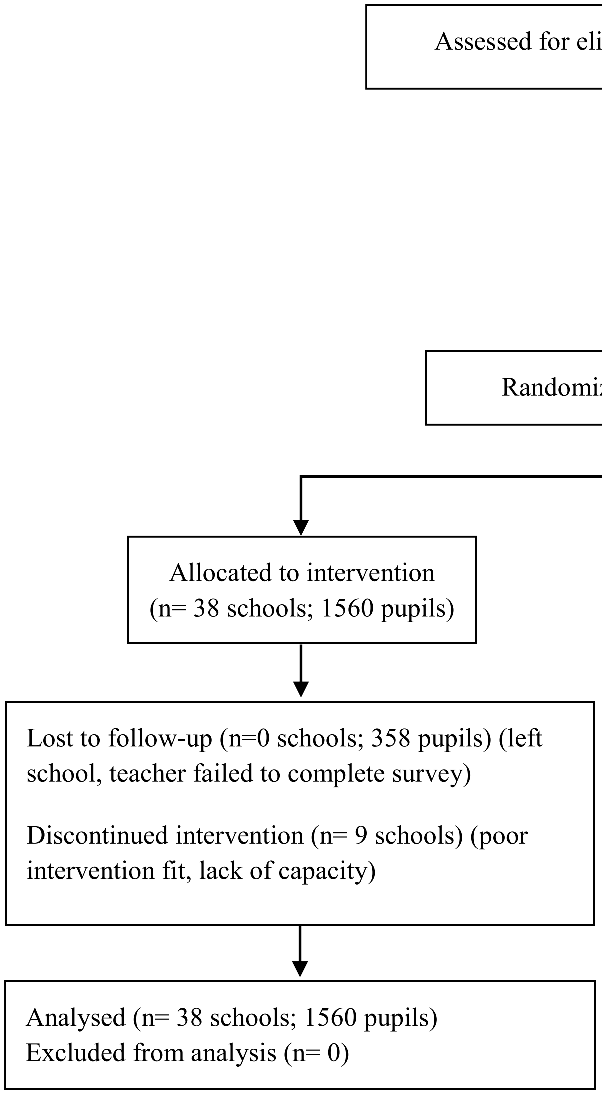
Skinner, B. F. (1945). The operational analysis of psychological terms. *Psychological Review*, *52*(5), 270–277. https://doi.org/10.1037/h0062535

Smith, S., Barajas, K., Ellis, B., Moore, C., McCauley, S., & Reichow, B. (2019). A meta-analytic review of randomized controlled trials of the Good Behavior Game. *Behavior Modification*, Online First. https://doi.org/10.1177/0145445519878670

Tanner-Smith, E. E., Durlak, J. A., & Marx, R. A. (2018). Empirically based mean effect size distributions for universal prevention programs targeting school-aged youth: A review of meta-analyses. *Prevention Science*, *19*(8), 1091–1101. https://doi.org/10.1007/s11121-018-0942-1

Tingstrom, D. H., Sterling-Turner, H. E., & Wilczynski, S. M. (2006). The Good Behavior Game: 1969-2002. *Behavior Modification*, *30*(2), 225–253. https://doi.org/10.1177/0145445503261165

Figure


Assessed for eligibility (n=203 schools)
Excluded (n=126 schools)
Not meeting inclusion criteria (n=2 schools)
Declined to participate (n=124 schools)
Randomized (n=77 schools)
Allocated to intervention (n= 38 schools; 1560 pupils)
Allocated to control (n= 39 schools; 1524 pupils)
Lost to follow-up (n=0 schools; 358 pupils) (left school, teacher failed to complete survey)
Discontinued intervention (n= 9 schools) (poor intervention fit, lack of capacity)
Lost to follow-up (n=0 schools; 214 pupils) (left school, teacher failed to complete survey)
Analysed (n= 38 schools; 1560 pupils)
Excluded from analysis (n= 0)
Analysed (n= 39 schools; 1524 pupils)
Excluded from analysis (n= 0)

Table

Table 1.

Demographic and descriptive data.

	School-level sample			Child-level sample		
	Overall (*N*=77)	GBG (*n*=38)	UP (*n*=39)	Overall (*N*=3,084)	GBG (*n*=1,560)	UP (*n*=1,524)
Demographics						
Size - FTE students on roll	306.9	298.2	315.4	-	-	-
Sex - % males	-	-	-	52.6%	50.4%	54.9%
Attendance - % days absence	4.2%	4.3%	4.2%	-	-	-
FSM - % eligible for FSM	26.0%	27.6%	24.5%	24.8%	27.4%	22.8%
Ethnicity - % White British	67.2%	67.6%	66.7%	65.8%		
EAL - % speaking EAL	22.6%	22.0%	23.2%	27.8%	26.1%	29.5%
SEND - % with SEND	19.5%	20.9%	18.2%	20.6%	23.1%	18.0%
Attainment - % achieving level 4+ in English and maths	75.5%	76.2%	74.9%	-	-	-
Outcomes	Min-Max		Mean		*SD*	
	GBG	UP	GBG	UP	GBG	UP
Disruptive behavior (baseline)	1-5.78	1-5.78	1.71	1.61	0.81	0.81
Disruptive behavior (follow-up)	1-5.67	1-6.00	1.74	1.65	0.86	0.84

Note. FTE = full time equivalent; FSM = free school meals; EAL = English as additional language; SEND = special educational needs and disabilities; GBG = Good Behavior Game; UP = usual practice; SD = standard deviation

Table 2.

Dosage data for GBG schools.

Dosage (GBG schools)	Min-Max	Mean	*SD*
Games a week (2015/16)[1]	0-4.45	1.96	1.14
Games a week (2016/17)	0-4.38	1.22	1.08
Minutes a week (2015/16)	0-64.25	27.21	17.60
Minutes a week (2016/17)	0-80.86	18.08	18.60
Dosage in minutes (2015/16)	0-1285	530.10	357.90
Dosage in minutes (2016/17)	0-2345	524.42	539.48
Total dosage in minutes	0-3535	1066.00	719.50

Note. GBG = Good Behavior Game; SD = standard deviation

[1] Game delivery delayed at 2015/16 due to initial training and scoreboard development, and included 20 weeks total delivery compared to 29 weeks total delivery in 2016/17.

Table 3.

Intent to treat and sub-group analyses (N = 3,084)

		Risk groups β (SE) [b (SE)]		
	Full sample	Others vs. Low	Others vs. Moderate	Others vs. High
School				
School size	.16 (.10)	.16 (.10)	.15 (.10)	.15 (.10)
% eligible for free school meals	.06 (.12)	.06 (.12)	.05 (.12)	.05 (.12)
% speaking English as additional language	-.19 (.16)	-.19 (.16)	-.20 (.16)	-.16 (.17)
ITT effects (if GBG)	**.22 (.25)**	**.23 (.31)**	**.12 (.26)**	**.27 (.25)**
		[.06 (.08)]	**[.03 (.07)]**	**[.07 (.07)]**
	d = .09	**d = .09**	**d = .05**	**d = .11**
Child				
Sex (if male)	.07 (.02)***	.06 (.02)**	.06 (.02)**	.07 (.02)***
Free school meals (if eligible)	.04 (.02)**	.03 (.02)	.04 (.02)	.04 (.02)
Special educational needs (if SEN)	.02 (.02)	.02 (.03)	.02 (.03)	.02 (.02)
Baseline concentration problems	.14 (.03)***	.14 (.03)***	.14 (.03)***	.14 (.03)***
Baseline disruptive behavior	.64 (.03)***	.64 (.03)***	.64 (.03)***	.64 (.03)***
Baseline pro-social behavior	.02 (.03)	.02 (.03)	.02 (.03)	.02 (.03)
CRE group (if at risk)		**-.00 (.04)**	**-.01 (.03)**	**.03 (.03)**
		[-.02 (.06)]	**[-.01 (.05)]**	**[.13 (.13)]**
Cross level Interactions				
CRE*Trial Group		**-.01 (.03)**	**.03 (.04)**	**-.05 (.03)**
		[-.01 (.07)]	**[.06 (.07)]**	**[-.23 (.15)]**

Note. CRE = cumulative risk exposure; GBG = Good Behavior Game; SE = standard error; ITT = intent-to-treat; SEN = special educational needs. Standardized estimates are reported. Unstandardized estimates in [] are also reported for the explanatory variables and interaction effects. In bold are ITT and interaction effects. * $p < .05$, ** $p < .01$, *** $p < .001$.

Table 4.

CACE moderate and high compliance predicting disruptive behavior (N = 3,084)

	CACE moderate compliance β (SE)		CACE high compliance β (SE)	
	Compliers (31%)	Non-compliers (69%)	Compliers (17%)	Non-compliers (83%)
School				
School size	.38 (.15)*	.41 (.07)***	.05 (.14)	.23 (.05)***
% eligible for free school meals	.04 (.19)	-.09 (.19)	.28 (.13)*	.20 (.09)*
% speaking English as additional language	-.34 (.17)*	-.30 (.10)**	-.09 (.11)	-.24 (.08)**
CACE effects (if GBG)	**-1.72 (.17)*** d = -1.35**	**-**	**-1.75 (.15)*** d = -1.14**	**-**
Child				
Gender (if male)	.05 (.04)	.07 (.03)**	.02 (.06)	.08 (.02)***
Free school meals (if eligible)	.07 (.04)	.04 (.03)	.08 (.06)	.05 (.02)*
Special educational needs (if with SEN)	-.00 (.04)	-.02 (.03)	.02 (.06)	-.01 (.03)
Baseline concentration problems	.19 (.08)*	.13 (.03)***	.30 (.08)***	.11 (.03)***
Baseline disruptive behavior	.60 (.05)***	.67 (.03)***	.57 (.07)***	.67 (.03)***
Baseline pro-social behavior	-.01 (.05)	.03 (.04)	-.06 (.08)	.03 (.04)
Entropy	.86		.85	

Note. CACE = complier average causal effect; GBG = Good Behavior Game; SEN = special educational needs; SE = standard error. Standardized estimates are reported. In bold are CACE effects. * $p < .05$, ** $p < .01$, *** $p < .001$.

Table 5.

CACE moderate compliance and sub-group analyses (N = 2,677)

	Risk groups β (SE) [b (SE)]					
	Others vs. Low		Others vs. Moderate		Others vs. High	
School						
School size	.31(.08)***	.27 (.15)	.39 (.13)**	.08 (.06)***	.27 (.16)	.25 (.10)*
% eligible for fee school meals	.06 (.08)	.13 (.14)	-.02 (.11)	.14 (.14)***	-.01 (.13)	.16 (.14)
% English as additional language	-.20 (.07)**	-.24 (.16)	-.38 (.12)**	-.14 (.14)	-.25 (.10)*	-.20 (.15)
CACE effects (if GBG)	**-1.84 (.15)*** [-1.34 (.14)***] d = -1.94**	-	**-1.65 (.17)*** [-.93 (.18)***] d = -1.31**	-	**-1.75 (.14)*** [-.89 (.18)***] d = -1.27**	-
Child						
Sex (if male)	-.01 (.08)	.16 (.04)***	.07 (.08)	.16 (.04)***	.12 (.07)	.16 (.04)***
Free school meals (if eligible)	.07 (.08)	.08 (.06)	.13 (.08)	.07 (.06)	.16 (.07)*	.08 (.06)
Special educational needs (if SEN)	-.13 (.09)	-.04 (.07)	-.04 (.09)	-.04 (.07)	.02 (.08)	-.04 (.07)
Baseline concentration problems	.21 (.05)***	.13 (.03)***	.20 (.06)***	.12 (.03)***	.19 (.06)**	.12 (.03)***
Baseline disruptive behavior	.52 (.06)***	.61 (.03)***	.53 (.06)***	.62 (.03)***	.56 (.06)***	.62 (.03)***
Baseline pro-social behavior	-.03 (.05)	.03 (.04)	-.02 (.05)	.03 (.04)	.01 (.05)	.03 (.04)
CRE group (if at risk)	**-1.04 (.16)**[-1.02 (.17)***]**	-	**.28 (.22) [.28 (.22)]**	-	**.81 (.34)* [.81 (.34)***	-
Cross level Interactions						
CRE*Trial Group	**.41 (.06)*** [.83 (.14)***]**	-	**-.11 (.10) [-.23 (.20)]**	-	**-.24 (.07)*** [-1.21 (.36)**]**	-
Entropy	.89		.90		.89	

Note. CRE = cumulative risk exposure; GBG = Good Behavior Game; SEN = special educational needs; SE = standard error. Standardized estimates are reported. Unstandardized estimates in [] are also reported for the explanatory variables and interaction effects. In bold are CACE and interaction effects. * $p < .05$, ** $p < .01$, *** $p < .001$.