'Good Behaviour Game' trial protocol

Universal school-based prevention: examining the impact of the Good Behaviour Game on health-related outcomes for children

1.1 Background

Mental health problems, defined as changes in thinking, mood and/or behaviour that impair functioning [1], will yield the highest disease burden in high-income countries by 2030, accounting for nearly 10% of disability-adjusted-life-years (DALYs) [2]. Global epidemiological data suggests that (i) up to 20% of children and adolescents experience clinically significant difficulties; (ii) suicide is the third leading cause of death for adolescents; and, (iii) 50% of adult mental health problems originate in childhood and adolescence [3]. The individual and societal impacts of such problems are huge, and include reduced quality of life, lost economic productivity, destabilisation of communities, and higher rates of health, education and social care utilisation [3]. In financial terms the cost is estimated to be over £105 billion annually in England [4]. This is a public health crisis requiring a co-ordinated, evidence-informed, multidisciplinary response.

Externalising problems (such as conduct disorder) are particularly noteworthy in the context of this project. Children with conduct disorder are more than 16 times more likely to be excluded from school than their peers [5], at a cost of £65,000 each [6]. When followed up in adulthood, those exhibiting difficulties at age 6-7 incur a two- (health and social care) to three-fold (criminal justice) increase in public sector costs [7]. Boys are particularly at risk, with those exhibiting conduct problems at age 10 significantly less likely to be employed at age 30 than their peers [8].

At a broader level, recent research on developmental cascades has yielded insights into how functioning in different domains (e.g. health, education) are developmentally related [9,10]. For example, the 'adjustment erosion' hypothesis – which posits that nascent externalising problems serve to undermine later academic achievement via their impact on relationships with teachers and peers – has received tentative empirical support [11]. Developing knowledge and understanding of the nature, magnitude and stability of such relationships over time provides an important contribution to developmental theory, which can in turn inform prevention and intervention. As Masten et al. [9] observed, it is, "critically important to study the processes, timing and conditions of spreading and amplifying effects and to learn when to do what to interrupt negative progressions… many of the best interventions studies of our time were designed with such considerations in mind" (p.742).

1.2 Universal school-based prevention

Schools are increasingly expected to play a central role in addressing the above problems and promoting well-being, with good reason. They play a central role in the lives of children and their families, and their reach is unparalleled [12]. Drawing on the inoculation metaphor that, "an ounce of prevention is worth a pound of cure", a universal approach to intervention has become a defining characteristic of education policy and practice in this area. However, there has been a distinct lack of rigorous, UK-based research on the efficacy of universal school-based preventive interventions in improving children's mental health. Reviews by NICE [13,14] and others [15–17] have highlighted that the overwhelming majority of the evidence base originates in the United States. Although this situation is beginning to change with numerous trials on-going in the UK, overall we still know very little about what interventions work, how and why they work, and for whom and under what conditions they work [18].

With specific reference to conduct disorders, this gap in the evidence base was recognised by NICE [19], who highlighted a need for research on, "interventions to prevent or treat conduct disorders [that] have been specially designed for delivery in the classroom" (p.40). They proposed a large-scale randomized controlled trial (RCT) design of at least 24-months duration, comparing the intervention to usual practice, reporting immediate and longer-term outcomes, the inclusion of an economic analysis, and investigation of response

moderators (e.g. implementation variability). It is on the basis of these principles that the trial described herein has been designed.

Our project is also timely given the broader context of concerns about children's behaviour and mental health. For example, a recent report based by the schools inspectorate Ofsted suggested that children lose up to an hour of learning each day as a direct consequence of disruptive behaviour in the classroom [(20)]. Unsurprisingly, the Department for Education has made this a policy priority, releasing guidance to schools on behaviour and mental health [(21)] and announcing a new school-based strategy to improve children's mental health care [(22)]. Finally, the proposed developmental links between health and educational outcomes noted in section 1.1 have been recognized by organisations such as Public Health England in their new briefing for educators [(23)].

1.3 The Good Behaviour Game

The Good Behaviour Game (hereafter referred to as GBG) is one of the most popular behaviour management systems for primary-aged children. It works at the level of the classroom social environment and has a strong theoretical base. The GBG also has extensive evidence supporting its use. Since its initial development over 40 years ago [(24)] multiple trials across the United States, the Netherlands and Belgium have attested to its efficacy in promoting a range of positive outcomes (e.g. increased pro-social behaviour, reduced substance abuse, aggression and criminality) [(25)]. A pilot study found the GBG to be both acceptable and feasible in English schools [(26)], setting the stage for our trial. Finally, a meta-analysis by Flower et al [(43)] indicated that the GBG is, "a highly effective intervention" (p.559), producing effects in the moderate to large range for outcomes that are directly relevant to this trial (e.g. aggressive behaviour). If such effects are replicated in our trial, they would be important from a public health perspective, particularly if they are maintained over time.

In the interests of clarity and transparency, we adapt the Template for Intervention Description and Replication (TIDieR) [(27)] to describe the GBG model:

1. Brief name

The Good Behaviour Game (GBG)

2. Why: Rationale, theory and/or goal of essential elements of the intervention

The GBG draws upon the principles of contingency management, in that children receive reinforcement when they engage in appropriate behaviours. However, the group-based orientation of the intervention means it also uses social learning theory, because pupils at-risk of developing conduct problems are able to learn from the appropriate behaviour being modelled by other team members. Finally, the GBG is informed by social field/life course theory, which posits that successful adaptation at different life stages is contingent upon an individual's ability to meet particular social task demands. In school, these task demands include being able to pay attention, work well with others, and obey rules. Success in social adaptation is rated both formally and informally by other members of the social field (e.g. teachers, peers). Social field theory predicts that improving the way in which teachers socialise children will improve their social adaptation. It is also predicted that early improvements in social adaptation will lead to better adaptation to other social fields later in life [(25)].

3. Who: recipients of the intervention

The GBG is a universal intervention and is therefore delivered to all children in a given class.

4. What (materials): Physical or informational materials used in the intervention

Participating schools receive GBG manuals that detail the programme theory, goals and procedures. Other materials include some tangible rewards (e.g. stickers), displays (e.g. scoreboard), and data forms for recording and monitoring purposes.

5. What (procedures): The procedures, activities and/or processes used in the intervention

The GBG is described by Tingstrom et al. [28] as an, "interdependent group-oriented contingency management procedure" (p. 225). Pupils in a given class are divided into mixed teams with up to 7 members. Strata can include a range of factors such as behaviour, academic ability, and gender. The teams then attempt to win the game as a means to access particular privileges/rewards. During the game period, the class teacher records the number of infractions to the following four rules among the teams: (1) We will work quietly, (2) We will be polite to others, (3) We will get out of seats with permission, and (4) We will follow directions. The game is 'won' by the team with the lowest number of infractions when it ends, although any team with fewer than four infractions also accesses the agreed reward [25,26]. Over the course of implementation of the GBG, there is a natural evolution in terms of the types of rewards used (from tangible rewards such as stickers to more abstract rewards such as free time), how long the game is played for (from 10 minutes to a whole lesson), at what frequency (from three times a week to every day), and when rewards are given (at the end of the game, end of the day, and at end of the week) [28,29]. Good behaviour achieved during the relatively brief 'game' periods is increasingly generalised to other activities and parts of the school day. Thus, the intervention leads to behaviour modification and intrinsic reinforcement so that modified behaviour is retained even after external reinforcement is removed (maintenance) and will be exhibited in all settings (generalization). These processes are documented by 'game' and 'probe' data collected by teachers during implementation [26].

6. Who (provider): Intervention implementers

The GBG is implemented by class teachers. Three days of training (2 days initial; 1 day follow-up) are by provided the American Institutes for Research (AIR) and Mentor UK. On-going technical support and assistance is provided by trained Mentor UK coaches. This comprises modelling, observation and feedback, ad-hoc email and telephone support, and provision of additional/booster training or information sessions as required. The GBG coaches are, in turn, supported by a Mentor UK lead coach and staff at AIR.

7. How: Mode of delivery

The GBG is implemented face-to-face during the normal school day. As it is a behavior management strategy rather than a taught curriculum, it requires no space in the class timetable.

8. Where: The location of the intervention

The GBG is implemented on-site in participating schools

9. When and how much: Duration and dosage of the intervention

In the context of the current trial, the GBG is delivered over two school years. As noted above, dosage evolves throughout the period of implementation in terms of both the duration of the game (from 10 minutes to a whole lesson), and the frequency at which it is played (from three times a week to every day).

10. Tailoring: Adaptation of the intervention

The GBG is a manualised intervention and participating teachers receive technical support and assistance in order to optimize fidelity of implementation. However, it is now accepted that some form of adaptation is inevitable and indeed may be desirable in order to improve local ownership and fit to context [30,31]. An important aspect of the GBG coach role is to support teachers to make adaptations that are in keeping with the goals and theory of the intervention [32].

1.4 The GBG education-related trial

Our research team secured funding from the Education Endowment Foundation (EEF) to conduct a 2-year cluster-randomised trial of the GBG in English primary schools that focuses on educational outcomes. All of the intervention costs (including technical support and assistance) have been being funded by the EEF. 77 schools were randomly assigned to deliver the GBG or continue usual practice with children aged 7-8 (N=3,085) starting in September 2015. To mitigate the risk of differential attrition [33] in the usual practice

group[1], these schools received £1,500 (single form entry schools; pro-rata by size) as an incentive for their continued participation, and subject to continued compliance with the data collection protocol.

The primary outcome of the EEF education-related trial is *children's attainment in reading* (Hodder Group Reading Test [34]). Secondary outcome measures are *children's behaviour* (Teacher Observation of Children's Adaptation checklist [35]) and *teachers' self-efficacy in classroom management* (Teachers' Sense of Efficacy Scale [36]), *stress* (Teacher Stress Inventory [37]) and *retention*. Teachers in all participating schools are also being surveyed regarding their usual practices in behaviour management (e.g. existing use of contingency management strategies) and provision of both universal and targeted interventions in related areas (e.g. social and emotional aspects of learning - SEAL). In the case of schools allocated to the intervention arm, this will allow us to determine the extent of programme differentiation (see section 2.4). In schools allocated to the control arm, it will allow us to describe in detail what 'usual practice' actually entails – a critical consideration for the trial.

The education-related trial includes a comprehensive implementation and process evaluation (IPE). We are assessing implementation dimensions such as fidelity, dosage, and quality via independent observations. Our assessment of process comprises longitudinal case studies of six schools that are exploring issues of social validity, acceptability and feasibility of the GBG via interviews, focus groups, observations and document analysis, drawing upon the perspective of multiple informants (e.g. children, teachers, parents, GBG coaches) (see section 2.4 for further details).

1.5 Aims and objectives

We sought funding to augment the education-related trial so as to (i) collect additional data on health-related outcomes, (ii) perform an economic evaluation, and (iii) assess longer-term outcomes. Our primary aim is to assess the efficacy of the GBG in improving mental health outcomes for children in English primary schools. This aim will be achieved by addressing the following objectives:

1. **To determine the impact of the GBG on health-related outcomes for children**

H1: Children in primary schools implementing the GBG over a two-year period will demonstrate significantly better mental health; psychological wellbeing (H1a), conduct problems (H1b) and emotional symptoms (H1c), sources of resilience; peer and social support (H1d) and school environment (H1e), school attendance (H1f), and significantly lower rates of bullying (social acceptance) (H1g) and exclusion from school (H1h) when compared to those children attending control schools.

Our primary outcome is children's mental health – this is consistent with both the GBG logic model [26] and life course/social field theory [25]. Sources of resilience (peer and social support and school environment) are included to assess the extent to which intervention exposure increases children's ability to draw upon these when they experience adversity (consistent with life course/social field theory [25]). Social acceptance (bullying) is included as a proxy for improved social adaptational status and positive interactions among peers (as predicted by the GBG logic model [26]). Finally, school attendance and exclusions are included in order to assess the extent to which improvements in the aforementioned domains translate into measureable change in school outcomes relating to engagement and behaviour (as predicted by the GBG logic model [26]).

2. **To determine the differential effects of the GBG for boys at-risk of developing conduct disorders**

H2: Boys at-risk of developing conduct disorders (defined as scoring in the borderline or abnormal ranges of the conduct problems sub scale of the teacher-rated Strengths and Difficulties Questionnaire [38] at baseline) in primary schools implementing the GBG over a two- year period will demonstrate significantly better outcomes in mental health; psychological wellbeing (H2a), conduct problems (H2b) and emotional

[1] Despite this, we still anticipate *some* imbalance in loss to follow up between the trial arms. However, we are reassured by a recent review by Hewitt [33] that found, "no indication that attrition altered the results in favour of either the treatment or the control" (p.1264) in a convenience sample of trials.

symptoms (H2c), sources of resilience; peer and social support (H2d) and school environment (H2e), school attendance (H2f), and significantly lower rates of bullying (social acceptance) (H2g) and exclusion from school (H2h) when compared to those at-risk boys attending control schools.

We expect amplified effects of the GBG for boys at-risk of developing conduct disorders on the basis of previous research findings [38]. The GBG procedure is likely to appeal particularly to boys given the gendered socialization of competitiveness [39]. Additionally, the gender ratio for the development of conduct disorders in childhood is approximately 3:1 in favour of boys [5].

3. **To determine the moderating influence of implementation variability on health-related outcomes in the GBG**

H3: Variation in implementation specifically, fidelity (H3a); dosage (H3b), quality (H3c), participant responsiveness (H3d), and reach (H3e), will moderate health-related outcomes in schools implementing the GBG.

Research across multiple disciplines has consistently demonstrated that interventions are rarely implemented as designed and, crucially, that variability in implementation is associated with variability in the achievement of expected outcomes [40]. The GBG is no exception [41,42]. Assessment of implementation is therefore a fundamental consideration in a trial of this nature. On the basis of implementation theory [43] and the programme's logic model [26], we expect the (i) the magnitude of the influence of implementation variability to be greater in our primary, proximal outcomes than in our secondary, distal outcomes, and (ii) that quality and participant responsiveness will moderate the relationship between implementation fidelity and outcomes. These proposed dimensions of implementation will be subject to factor analyses to explore their independence ahead of the main analyses pertaining to H3.

4. **To determine the sustainability of the GBG's effects on health- and education-related outcomes**

H4: The effects of GBG on mental health; psychological wellbeing (H4a), conduct problems (H4b) and emotional symptoms (H4c), sources of resilience; peer and social support (H4d) and school environment (H4e), school attendance (H4f), and significantly lower rates of bullying (social acceptance) (H4g) and exclusion from school (H4h), and improvements in reading attainment (H4i) and pro-social behaviour (H4j), and reductions in concentration problems (H4k) and disruptive behaviour (H4l), will be maintained at 12 and 24-month post-intervention follow-up .

This hypothesis is based upon existing evidence of the sustained effects of the GBG [25], and life course/social field theory, which suggests that effective socialization of behaviour can yield a lasting influence on children's social adaptational status [38]. It is critical to include an 'interim' (e.g. 12 month) follow-up so that we are able to model the maintenance of intervention effects with greater precision.

5. **To determine the nature and magnitude of developmental cascades between children's educational and health-related outcomes over time**

H5: Children's educational and health-related outcomes will be related over time.

We make multiple predictions here, drawing upon developmental cascades [10] and ecological systems [44] theories. First, we test the adjustment erosion hypothesis – namely, that early conduct problems (H5a), emotional symptoms (H5b) and experience of bullying (social acceptance) (H5c) will be negatively associated with later academic achievement. Second, we test the academic incompetence hypothesis - that nascent academic difficulties will serve as a trigger for later mental health problems (conduct problems H5d; emotional symptoms H5e). Third, we test the shared risk (or common cause) hypothesis – that the cascading effects noted above are a function of 'third variables' affecting multiple and inter-related domains of development [11]. In this proposal we utilise socio-economic disadvantage (H5f) and special educational needs (H5g) as our risk markers. Finally, we assess the extent to which the sources of resilience, peer and social support and school environment, mitigate the effects of exposure to these risk markers.

6. **To assess the extent to which the GBG can be regarded as providing value for money**

H6: The GBG will represent an efficient use of resources when considered from the perspective of the UK Treasury, resulting in a social rate of return that is considered acceptable

There is good reason to propose that the GBG will prove to be cost-effective. It is financially viable[2], costing approximately £3,700 per class based on an optimal level of training and technical support and assistance[3], requires relatively little training for implementers (3 days) and does not require any curriculum time (meaning that important teaching and learning time is not displaced). Furthermore, the GBG has been demonstrated to be efficacious in improving outcomes that are directly relevant to increased QALYs [(45)].

2.1 Research Design

A 2-year cluster-randomised trial [(46)] with 2-year follow-up period is being utilised. Participating schools will are the unit of randomisation. The allocation procedure was conducted independently by the Manchester Academic Health Science Centre Clinical Trials Unit (UK CRC CTU 9). A minimisation algorithm was applied to the randomisation to ensure balance across the arms of the trial in terms of the proportion of children eligible for free school meals (FSM) and school size. Schools randomly allocated to the intervention arm are implementing the GBG (with technical support and assistance) for a period of two years. Schools randomly allocated to the control arm are continuing their usual practice during this period. As noted above, our study protocol enables us to document fully what 'usual practice' entailed. Based on experience we expect teachers to report using a variety of behaviour management strategies, some of which may be loosely based around contingency management principles. Importantly, these are typically not systematic in nature, are individually rather than group oriented, and often inadvertently draw attention to maladaptive behaviour (e.g. traffic light system with children being placed 'on red'). Use of a range of related interventions such as the primary version of the SEAL programme [(47)] was also expected and will be documented. We drew on existing research on teachers' use of different behaviour management strategies [(48)] to ensure that our approach was comprehensive.

The augmentation of health-related outcomes to the existing education-related trial uses a post-test experimental design with 2-year follow-up. A post-test design confers numerous advantages, including reduced data burden for schools and no pretest sensitisation effects [(49)]. Furthermore, Gorard [(50)] argues that it is "generally at least as safe as its alternatives, and is sometimes preferable or more feasible than... pre-and-post-test designs" (p.2). The large number of schools in the trial means that random allocation should ensure that any pre-existing differences do not bias our results [(51)].

Hence, the education-related trial assessed educational outcomes at baseline in June/July 2015 (T1). Following interim assessment of outcomes in June/July 2016 (T2), the final post-test will be June/July 2017 (T3). It is at this point that we will also capture the health-related outcomes. We are then to conduct follow-up assessment of both educational- and health-related outcomes at 12- and 24-months respectively (T4 and T5). Assessment of implementation is being conducted during the main trial phase (e.g. between T1 and T3).

2.2 Setting and target population

The setting of the trial is primary schools in England. 'Primary schools' were restricted to mainstream institutions providing education for children from the ages of 4-11. 77 schools were recruited from the Greater Manchester region (e.g. Bury, Manchester, Salford, Stockport, and Wigan) in addition to conurbations in West Yorkshire (e.g. Bradford and Leeds), South Yorkshire (e.g. Sheffield and Barnsley), the West Midlands (e.g. Telford), and the East Midlands (e.g. Nottingham). These areas provide great diversity

[2] The cost per class is less than the additional funding each primary school in England receives annually for any 2-3 children eligible for the 'pupil premium' (e.g. those who are looked after and/or eligible for free school meals).

[3] NB: these are start-up costs; in subsequent years the cost per child drops significantly.

in their urbanicity, socio-economic status, ethnicity, and other relevant factors that will help to ensure that our research setting is representative of England.

Our target population is children aged 7-8 (Year 3) attending English primary schools. All children who were on a given school's full-time roll in Year 3 at the start of the 2015/16 school year and provided parental consent (opt-out procedure) were potential participants (N=3,085 pupils). A sub-group of particular interest within the study population is boys at-risk of developing conduct disorder (N=337, 11%) (see Hypothesis 2 in section 1.5 above).

2.3 Socio-economic position and inequalities

The study can make a substantial contribution to addressing established inequalities in mental health. By definition, universal school-based interventions can reach children with or at risk for developing difficulties who may not access support through usual care pathways [19]. Our distinction between primary (H1) and secondary (H2) intervention effects ensures that we will be able to fully analyse the extent to which the GBG is effective in addressing this inequity. As standard, our data collection protocol allows us to take into account the socio-economic position of participants (see 2.5). Indeed, given the well-established links between socio-economic disadvantage and the development of mental health difficulties in childhood [5], inclusion of free school meal eligibility and/or Index of Deprivation Affecting Children data as a co-variate in our statistical models is a fundamental consideration. Our developmental cascade modelling (H5 in section 1.5) will further our understanding of the effects of socio-economic position on developmental trajectories and the potential mitigating effects of sources of resilience in peer and social support and the school environment. Finally, our assessment of process (see section 2.4) will allow us to generate explanatory data on the mechanisms through which the GBG enables positive change among marginalised groups.

In terms of our research procedures, we will strive to ensure equitable access by, for example, providing on-site support for participants whose special educational needs make it difficult for them to complete study measures (e.g. those with dyslexia). All measures and study documentation (e.g. information sheets) will be screened by our Trial Steering Committee (TSC) (in particular, young research advisors from Common Room - see section 3.2) to ensure that the language and presentation format used is accessible and engaging and that technical language and jargon are avoided.

2.4 Implementation and process evaluation

As noted in section 1.4, the IPE is being funded as part of the education-related trial. As we plan to use the IPE data in the health-related trial (e.g. H3 in section 1.5), we provide details of our protocol here. Our assessment of implementation will seek to determine the extent of variability in *fidelity* (to what extent do teachers adhere to the GBG manual?), *dosage* (how frequently is the GBG played and for how long?), *quality* (how well do teachers deliver the components of the GBG?), *participant responsiveness* (do children engage with the GBG?) and *reach* (what is the rate and scope of participation in the GBG across the class?). Our data collection protocol was informed by those used in previous GBG studies [29], our own work in other trials (e.g. PATHS), and naturally occurring data (e.g. game and probe data relating to rule infraction can be used as a proxy for participant responsiveness [26]). Data is being generated through annual independent, structured observations. The structured observations were piloted and refined ahead of the main trial to establish inter-rater reliability and ensure that the measure is fit for purpose. Additional video footage of GBG implementation was then used in order to generate inter-rater reliability data for each indicator. These analyses demonstrated exceptionally good IRR, e.g. Cohen's Kappa for our nominal procedural fidelity items is 0.95, indicative of near perfect agreement.

Each class in the GBG arm of the trial is observed twice – once in 2015/16, and once in 2016/17. In line with the evidence that the mean ratings from two time-points will be more strongly associated with outcomes than a single time-point [53, 54], we will aggregate this data prior to analysis.

Our assessment of process comprises case studies of six GBG schools representative of the school sample. The case studies are exploring issues of social validity, acceptability and feasibility of the GBG. Here we

drew upon the recent GBG UK pilot [26], related studies of school-based interventions [52] and adapted existing rubrics from the implementation literature [53] to inform our data generation. We are also using the case studies to explore a range of factors affecting implementation at different levels and domains identified in the literature [30,54,55]. A multi-method (e.g. interviews, focus groups, observations), multi-informant (e.g. children, teachers, parents, GBG coaches) approach is being utilised. We will draw on the recent review of theories in implementation science [59] to explore and understand our findings.

2.5 Assessment of outcomes

Instrumentation criteria

In selecting our primary and secondary outcomes measures we have used the following criteria: (i) goodness-of-fit with study parameters (e.g. age of participants, domains of interest), (ii) psychometric properties (using the thresholds set by Terwee [56]), (iii) brevity and accessibility, and (iv) use in similar or related research published in peer-reviewed journals. The following measures have been approved by the TSC, including Common Room, who are our PPI experts (see section 3.2; NB: some measures from original proposal have been substituted following TSC discussion to improve accessibility and/or reduce data burden – see appendix for summary).

Primary outcome measures

The KIDSCREEN27 Psychological Wellbeing domain provides self-reported assessment of mental health among children [57]. It is brief, comprising 7 items in which respondents read a statement (an example is, "thinking about last week, have you been in a good mood") and indicate their agreement on a 5-point scale (never, seldom, quite often, very often, always). The KIDSCREEN27 was designed and validated for use with children aged 8 and above. Finally, it is psychometrically sound, with good internal consistency (alpha co-efficient of 0.84), a robust factor structure (established through CFA), and strong predictive validity (for example, discriminates between those identified with mental health problems as assessed by the Strengths & Difficulties Questionnaire, ES = 0.68, and correlates with other similar measures such as, the Youth Quality of Life Instrument (.63), Child Health questionnaire (.36) and Child Health and Illness Profile (.62) (57). The measure has been used in a previous randomised trial of a school-based mental health intervention [58].

The Psychological Wellbeing outcome data will be supplemented with the emotional symptoms and conduct problems subscales of the teacher informant-report version of the SDQ [38]. This will allow for multi-informant triangulation of our two primary mental health outcomes – internalising and externalising difficulties – thereby increasing the validity of the study with limited additional data burden for schools/teachers. The SDQ subscales in question comprise a total of 10 items in which respondents read a statement (an example is, "Often has temper tantrums or hot tempers") and indicate their agreement on a 3-point scale (not true, somewhat true, certainly true). The SDQ is psychometrically robust and clinically valid [38].

Secondary outcome measures

Health-related quality of life

The Child Health Utility 9D (CHU9D) [59] is a measure of paediatric health-related quality of life that can be used in economic evaluation. It is very brief, comprising only 9 items (one for each health-state dimension, e.g. "sad") with 5 response options for each (e.g. "I don't feel sad today/I feel a little bit sad today/I feel a bit sad today/I feel quite sad today/I feel very sad today"). It was designed specifically for use with children aged 7 and above and has a very low reading age (Flesh Kincaid index = c.2.4). The CHU9D has been used in a number of related studies, including our on-going PATHS trial. It is psychometrically sound, with demonstrable discriminative, construct and convergent validity [60,61]. The development of the CHU9D included standard gamble modeling of preference weights for health states defined by the instrument, enabling the calculation of QALYs [62].

Bullying

The KIDSCREEN52 Social Acceptance domain provides self-reported assessment of experience of bullying among children [57]. It is brief, comprising 3 items in which respondents read a statement (an example is, "thinking about last week, have other girls and boys made fun of you") and indicate their agreement on a 5-point scale (never, seldom, quite often, very often, always). The KIDSCREEN52 was designed and validated for use with children aged 8 and above. Finally, it is psychometrically sound, with good internal consistency (alpha co-efficient of 0.77), a robust factor structure (established through CFA), and strong predictive validity (for example, discriminates between those identified with mental health problems as assessed by the Strengths & Difficulties Questionnaire, ES = 0.70) [57].

Resilience

The KIDSCREEN27 Social Support and Peers, and School Environment domains provides self-reported assessment of resilience among children [57]. It is brief, comprising 4 items in each domain in which respondents read a statement (an example is, "thinking about last week, have you spend time with your friends (social support an peers), and "thinking about last week, have you been happy at school" (school environment)") and indicate their agreement on a 5-point scale (never, seldom, quite often, very often, always (social support and peers), and not at all, slightly, moderately, very, extremely (school environment)). The KIDSCREEN27 was designed and validated for use with children aged 8 and above. Finally, it is psychometrically sound, with good internal consistency (alpha co-efficients of 0.81 for each), a robust factor structure (established through CFA), and strong predictive validity (for example, discriminates between those identified with mental health problems as assessed by the Strengths & Difficulties Questionnaire, ES = 0.5 (social support and peers) and 0.62 (school environment) [57]. The measures have been used in a previous randomised trial of a school-based mental health intervention [58].

School attendance and exclusions

Data on children's school attendance (recorded as % half-days absent) and exclusions (fixed-term and permanent) will be extracted from the National Pupil Database (NPD) (see 2.6 below).

Reading

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

Behaviour

Children's behaviour will be assessed using the Teacher Observation of Children's Adaptation checklist (TOCA-C; [35]). This 21-item scale provides indices of children's concentration problems and disruptive behaviour and pro-social behaviour. Raters read statements about a child (e.g. "Pays attention") and endorse them on a 6-point scale (Never/Rarely/Sometimes/Often/Very Often/Almost Always).

At baseline we employed the teacher-rated conduct problems subscale of the Strengths and Difficulties Questionnaire (SDQ [63]) in order to identify our at-risk sample. This 5-item scale requires raters to read statements about a child's behaviour (e.g. "Often has temper tantrums or hot tempers") and endorse them on a 3-point scale (Not True/Somewhat True/Certainly True). The subscale produces a score of 0-10, with 0-2, 3 and 4-10 representing the normal, borderline and abnormal ranges respectively. At-risk status is defined as scoring in the borderline or abnormal range on this measure at T1.

2.6 Co-variates

As part of the education-related trial we will be collecting background data on both schools (e.g. size, proportion of children eligible for FSM, urbanicity) and children (e.g. sex, FSM eligibility, special educational needs) for use as co-variates in in our various analyses. School-level data is taken from DfE performance tables data and child-level data has been extracted from the NPD. The NPD also provides an anonymised

child reference number that we are using to ensure accurate data matching (e.g. across time and between informants).

2.7 Power and sample size

The education-related trial is powered to detect an effect size of 0.13 for reading in an intention-to-treat analysis, with a baseline sample (77 schools, N=3,085 children, average of 40 pupils per cluster) ICC of 0.06, an assumed pre-post correlation of 0.7 (based on EEF estimates), and Power and Alpha set to 0.8 and 0.05 respectively.

Thus, at post-test, and for the purposes of the health-related trial described herein, we estimate 71 schools and a sample of N=2,840 (inclusive of anticipated attrition; see CONSORT diagram in appendix). We assume an ICC of no more than 0.02 for our primary outcome measure of mental health. This is based upon the findings of a major secondary analysis of mental health data from the Avon Longitudinal Study of Parents and Children (64). With Power and Alpha set at 0.8 and 0.05 respectively, the minimum detectable effect size (MDES) will therefore be 0.14 in an intention-to-treat analysis. For our proposed sub-group analysis (see H2 in section 1.5), cluster size reduces to 4 (21% of the c. 1,420 boys in the sample assumed to meet 'at risk' criteria [5]) = 298.2, or 4.2 per cluster) giving us a MDES of 0.35.

2.8 Analytical strategy

We will construct a detailed statistical analysis plan under the supervision of Boehnke (Co-I). Our provisional analytical strategy is as follows:

H1: Children in primary schools implementing the GBG over a two-year period will demonstrate significantly better mental health; psychological wellbeing (H1a), conduct problems (H1b) and emotional symptoms (H1c), sources of resilience; peer and social support (H1d) and school environment (H1e), school attendance (H1f), and significantly lower rates of bullying (social acceptance) (H1g) and exclusion from school (H1h) when compared to those children attending control schools.

We will conduct intention-to-treat analysis [(66)] for the primary outcomes (mental health variables: psychological wellbeing, conduct problems and emotional symptoms) and secondary outcomes (peer and social support, school environment, school attendance, social acceptance, and exclusions). Multi-level modelling (MLM) fixed effects and random slopes, using MLWin. Two-level (school, pupil) hierarchical models will be fitted to account for nested nature of dataset, with the outcome variable as the response variable.

Initially, empty ('unconditional') models will be fitted, entering only the levels and no explanatory variables. This will allow approximations of the proportion of unexplained variance attributable to each level of the model. Subsequently, a full ('conditional') model will be fitted, to include minimisation variables (% FSM and school size) and trial group (GBG vs. control) at the school level, and gender, FSM eligibility (given their associations with the response variable) at the child level. An intervention effect will be noted if the co-efficient associated with the trial group variable is statistically significant and in the expected direction.

Data will be standardised prior to modeling, facilitating comparison of effect sizes within and across models. In determining the importance of any intervention effects, we will go beyond standard (and, arguably, arbitrary) considerations of effect size classification [(67)], instead utilising Hill and colleagues' [(68)] three empirical benchmarks for practical significance (normative growth, policy-relevant gaps, effects of similar interventions)

H2: Boys at-risk of developing conduct disorders (defined as scoring in the borderline or abnormal ranges of the conduct problems sub scale of the teacher-rated Strengths and Difficulties Questionnaire [(38)] at baseline) in primary schools implementing the GBG over a two- year period will demonstrate significantly better outcomes in mental health; psychological wellbeing (H2a), conduct problems (H2b) and emotional symptoms (H2c), sources of resilience; peer and social support (H2d) and school environment (H2e), school

attendance (H2f), and significantly lower rates of bullying (social acceptance) (H2g) and exclusion from school (H2h) when compared to those at-risk boys attending control schools.

The above models will be extended to include baseline risk-status at the child level (e.g. normal vs. borderline/abnormal) such that two-way cross-level interactions (intervention group*risk status*gender) can be examined[4].

H3: Variation in implementation specifically, fidelity (H3a); dosage (H3b), quality (H3c), participant responsiveness (H3d), and reach (H3e), will moderate health-related outcomes in schools implementing the GBG.

First, in order to streamline analyses and thus reduce the likelihood of 'model overfitting', avoid collinearity, and establish clear differentiation between implementation constructs, the observer-rated implementation data, derived from the education-related trial, pertaining to fidelity, quality, responsiveness and reach will subjected to exploratory factor analysis. The resultant factors will then be modeled alongside the dosage (cumulative intervention intensity) and reach (proportion of class present).

We will then construct two-level hierarchical linear models (child, class/teacher) of data from GBG schools only, with implementation entered at the class/teacher and school levels. Following Peugh [(69)] we will report both global (proportional reduction in variance in the response variable associated with the inclusion of all explanatory variables) and local (the influence of individual variables on the response variable) effects for these models. In the case of the former, this will allow us to test our prediction that the magnitude of influence of implementation variability will be greater in primary, proximal outcomes than in secondary, distal outcomes. In the case of the latter, it will allow us to determine the both the impact of the GBG and the influence of individual aspects of implementation (e.g. fidelity) after controlling for a range of co-variates.

H4: The effects of GBG on mental health; psychological wellbeing (H4a), conduct problems (H4b) and emotional symptoms (H4c), sources of resilience; peer and social support (H4d) and school environment (H4e), school attendance (H4f), and significantly lower rates of bullying (social acceptance) (H4g) and exclusion from school (H4h), and improvements in reading attainment (H4i) and pro-social behaviour (H4j), and reductions in concentration problems (H4k) and disruptive behaviour (H4l), will be maintained at 12 and 24-month post-intervention follow-up .

We will conduct intention-to-treat analysis for the primary outcomes (mental health variables: psychological wellbeing, conduct problems and emotional symptoms) and secondary outcomes (resilience, school attendance, bullying, and exclusions) and education variable (reading). Two analyses will be ran, i) outcomes at 12-month follow-up and outcomes at 24-month follow-up.

Two-level (school, pupil) hierarchical models will be fitted. Initially, empty ('unconditional') models will be fitted, entering only the levels and no explanatory variables. This will allow approximations of the proportion of unexplained variance attributable to each level of the model. Subsequently, a full ('conditional') model will be fitted, to include minimisation variables (% FSM and school size) and trial group (GBG vs. control) at the school level, and gender, FSM eligibility (given their associations with the response variable). An intervention effect will be noted if the co-efficient associated with the trial group variable is statistically significant and in the expected direction.

H5: Children's educational and health-related outcomes will be related over time.

We will produce structural equation models of the longitudinal associations between our various educational and health-related outcomes. Critically, true developmental cascade models must model

[4] This approach can also be applied to assess the extent to which the GBG addresses inequalities by substituting the grouping variable at the child level (e.g. FSM eligible vs. not eligible for socio-economic position)

cross-domain associations among three areas of functioning (e.g. emotional symptoms, conduct problems and academic attainment) across at least *three* time points (e.g. T3, T4, T5) [(11)]. Following Moilanen [(11)] we will initially test two models, with and without shared risk, assessing comparative model fit and changes in the magnitude and statistical significance of individual path co-efficients. A third model that incorporates sources of resilience (e.g. peer and social support and school environment) as a moderator of the relationship between risk markers and outcome variables will then be tested.

H6: The GBG will represent an efficient use of resources when considered from the perspective of the UK Treasury, resulting in a social rate of return that is considered acceptable

An economic evaluation will be conducted from the perspective of the UK Treasury. Staff inputs and materials associated with GBG implementation will be documented following discussions with schools, AIR and Mentor UK and translated into costs using published unit costs. The agencies that incur costs will be clearly specified, as will the agencies that benefit from possible reductions in resource utilisation, so as to enable inter-sectoral comparisons. The overall costs of providing the programme, cost per school and cost per pupil will be computed, along with degree of variation among participants.

We will also measure the extent of resource utilisation of health services, educational support and social services by the participants (and families) prior to the GBG implementation, as a result of the intervention and at 1-year and 2-year follow-up. Changes in utilisation will be monetarised using published unit costs and merged with the costs of the GBG programme to generate net cost of programme. The net cost will be used in conjunction with the specified outcomes to produce a cost-consequences analysis. Cost-consequences analysis will be used as the outcomes from the intervention cannot be measured in the same units. This enables different decision makers to place their own weights on the different benefits and on costs. It will include an estimate of the cost per QALY gained (via the CHU9-D), while the potential QALY gains will be utilised in a net–benefit analysis based on accepted thresholds (e.g. NICE assessments) of 'value for money'. These will also enable a social rate of return on investment to be derived based on expenditure incurred on the GBG intervention.

A series of sensitivity analyses will determine the impact of parameter variation on baseline values and a probabilistic sensitivity analysis undertaken to assess the extent to which the GBG intervention can be regarded as representing an efficient use of public funds. GBG impact at post-test and follow-up will be modelled using a range of timescales (10 years, 20 years and lifetime durations).

3.1 Ethical considerations

The research outlined herein was subject to rigorous internal scrutiny and approved by our University Research Ethics Committee (UREC) (ref 15126), who will also provide on-going review should any issues arise during the course of the study. All members of the research team have full Disclosure and Barring Service (formerly Criminal Records Bureau) checks completed.

The design and conduct of the research is in accordance with the six key principles in the Economic and Social Research Council's (ESRC) research ethics framework: (i) *Integrity, quality and transparency* – this will be assured through multiple oversight and governance arrangements (e.g. NIHR, UREC, the TSC). We will also produce periodic public reports to facilitate an open and transparent process; (ii) *Informed consent* – schools, parents and their children have been provided with information sheets about the purpose, methods and intended uses of the research, what 'participation' entails, and any anticipated risks and benefits. Different versions have been produced that are appropriate to the different stakeholders. Additionally, a contact number has been provided to enable additional queries to answered. Consent was sought at three levels – school, parent and child (assent). In view of the large sample size and measurement protocols (see section 2), an opt-out procedure was used for parental consent[5]; (iii) *Confidentiality and anonymity* – all data will be kept anonymous and confidential except in the event of a child protection issue

[5] Except for in our assessment of process in case study schools, where we will use opt-in consent in view of the methods of data generation (e.g. child focus groups).

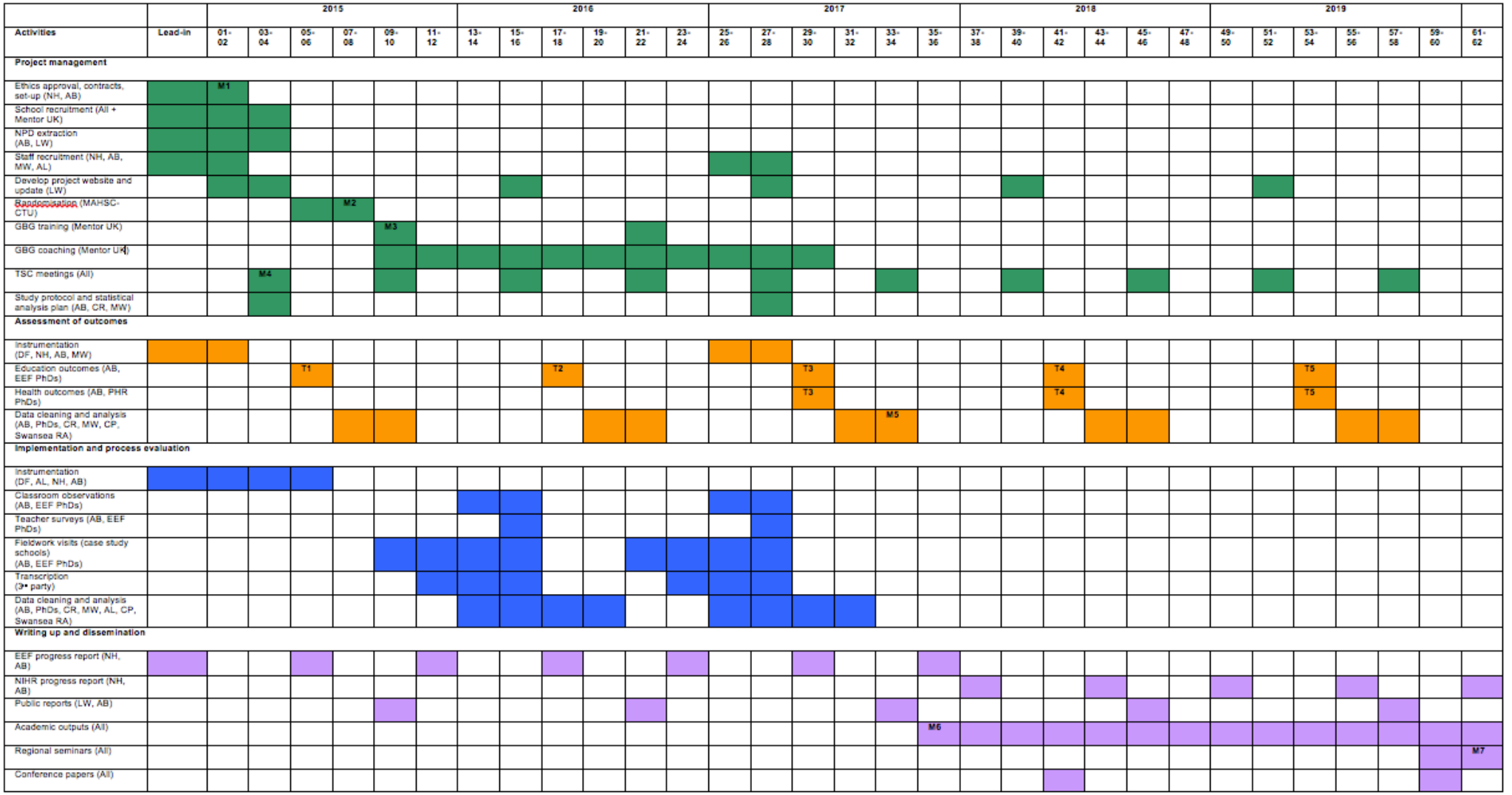
(for example, a child disclosing details of abuse to a researcher during a fieldwork visit), at which point standard safeguarding protocols will be followed. Data security for online surveys will be ensured using hypertext transfer protocol secure (HTTPS) data encryption. Data matching will be achieved through the triangulated use of name, date of birth and anonymised pupil matching reference numbers, but these will be removed prior to analysis. All data will be held safely on secure drives, with Microsoft Best Practice guidelines followed. Data will be held behind both internal and external firewalls, and physical transportation (e.g. on flash drives) will be prohibited; (iv) *Voluntary participation* – we will strive to ensure that all participants provide data on a completely voluntary basis. As the outcome measures are likely to be taken on a whole-class basis, we will encourage schools to provide alternative activities for children who do not wish to participate. The right to withdraw at any point of the study without needing to give a reason will be reinforced in duplicate (e.g. on information and consent sheets, in online survey instructions); (v) *Avoidance of harm* – the design of the research minimises the risk of harm to participants. As a failsafe, members of our research team will have reviewed participating schools' health and safety protocols and will act accordingly in the event of such an incident. In terms of emotional harm, in the event of a participant becoming upset or distressed at any point in the research, the researcher will immediately cease data collection and contact an a-priori nominated member of school staff to provide support. Preventive measures will also be in place – for example, contact details of organisations who can provide independent support and advice on social and emotional issues (e.g. Childline) will be made available to all participants; and (vi) *Independence* – the research team has no affiliation with or financial interests in the GBG, the American Institutes for Research (US) or Mentor (UK). We have no conflicts of interest and are able to conduct the research objectively and impartially.

3.2 Research governance

The University of Manchester is the sponsor of the research. Governance will be provided through our existing support systems (e.g. UREC, Research Development Officer, Finances and Accounts, and Research Strategy Committee). We are also be held to account by our TSC, which will meet every 6-12 months and comprises an independent academic chair, young research advisors from Common Room, an independent statistician (e.g. from MAHSC-CTU), a relevant independent voluntary sector representative (e.g. Young Minds), and representatives from the research team, Mentor, and the EEF. The TSC will help to ensure social validity by providing governance on the design and conduct of the study. For example, an early task carried out by Common Room was to assess our outcome measurement protocol (see section 2.5) to ensure that it was meaningful and accessible to children and young people. The TSC is the primary means through which we will ensure that Patient and Public Involvement requirements are met.

3.6. Gantt chart and milestones

In the interests of clarity we provide a Gantt chart for the whole trial, inclusive of the education- and health-related components:


2015
2016
2017
2018
2019
Activities
Lead-in
01-02
03-04
05-06
07-08
09-10
11-12
13-14
15-16
17-18
19-20
21-22
23-24
25-26
27-28
29-30
31-32
33-34
35-36
37-38
39-40
41-42
43-44
45-46
47-48
49-50
51-52
53-54
55-56
57-58
59-60
61-62
Project management
Ethics approval, contracts, set-up (NH, AB)
M1
School recruitment (All + Mentor UK)
NPD extraction (AB, LW)
Staff recruitment (NH, AB, MW, AL)
Develop project website and update (LW)
Rapporteuring (MAHSC-CTU)
M2
GBG training (Mentor UK)
M3
GBG coaching (Mentor UK)
TSC meetings (All)
M4
Study protocol and statistical analysis plan (AB, CR, MW)
Assessment of outcomes
Instrumentation (DF, NH, AB, MW)
Education outcomes (AB, EEF PhDs)
T1
T2
T3
T4
T5
Health outcomes (AB, PHR PhDs)
T3
T4
T5
Data cleaning and analysis (AB, PhDs, CR, MW, CP, Swansea RA)
M5
Implementation and process evaluation
Instrumentation (DF, AL, NH, AB)
Classroom observations (AB, EEF PhDs)
Teacher surveys (AB, EEF PhDs)
Fieldwork visits (case study schools) (AB, EEF PhDs)
Transcription (3rd party)
Data cleaning and analysis (AB, PhDs, CR, MW, AL, CP, Swansea RA)
Writing up and dissemination
EEF progress report (NH, AB)
NIHR progress report (NH, AB)
Public reports (LW, AB)
Academic outputs (All)
M6
Regional seminars (All)
M7
Conference papers (All)

3.7 References

1. Murphey D, Barry M, Vaughn B. Mental health disorders. Child Trends. 2013;(January):1–10.
2. Mathers CD, Loncar D. Projections of global mortality and burden of disease from 2002 to 2030. PLoS Med. 2006 Nov;3(11):e442.
3. Belfer ML. Child and adolescent mental disorders: the magnitude of the problem across the globe. J Child Psychol Psychiatry. 2008 Mar;49(3):226–36.
4. Centre for Mental Health. The economic and social costs of mental health problems in 2009/10. London; 2010.
5. Green H, McGinnity A, Meltzer H, Ford T, Goodman R. Mental Health of Children and Young People in Great Britain. Area. 2005.
6. Brookes M, Goodall E, Heady L. Misspent youth: the costs of truancy and exclusion – a guide for donors and funders. London; 2007.
7. D'Amico F, Knapp M, Beecham J, Sandberg S, Taylor E, Sayal K. Use of services and associated costs for young adults with childhood hyperactivity/conduct problems: 20-year follow-up. Br J Psychiatry [Internet]. 2014 Jun [cited 2014 Jul 23];204(6):441–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24676966
8. Knapp M, King D, Healey A, Thomas C. Economic outcomes in adulthood and their associations with antisocial conduct, attention deficit and anxiety problems in childhood. J Ment Health Policy Econ [Internet]. 2011 Sep [cited 2014 Jul 30];14(3):137–47. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22116171
9. Masten AS, Roisman GI, Long JD, Burt KB, Obradovic J, Riley JR, et al. Developmental cascades: linking academic achievement and externalizing and internalizing symptoms over 20 years. Dev Psychol [Internet]. 2005 Sep [cited 2012 Oct 25];41(5):733–46. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16173871
10. Masten AS, Cicchetti D. Developmental cascades. Dev Psychopathol [Internet]. 2010 Aug [cited 2012 Oct 25];22(3):491–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20576173
11. Moilanen KL, Shaw DS, Maxwell KL. Developmental cascades: externalizing, internalizing, and academic competence from middle childhood to early adolescence. Dev Psychopathol [Internet]. 2010 Aug [cited 2012 Oct 19];22(3):635–53. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3168570&tool=pmcentrez&rendertype=abstract
12. Greenberg MT. School-based prevention: current status and future challenges. Eff Educ. 2010;2:27–52.
13. Adi Y, Mcmillan AS, Kiloran A, Stewart-brown S, Medical W, Stewart-brown CS, et al. Systematic review of the effectiveness of interventions to promote mental wellbeing in primary schools Report 3 : Universal Approaches with focus on prevention of violence and bullying. Sci York. (September 2007):1–106.
14. Blank L, Baxter S, Goyder L, Guillaume L, Wilkinson A, S. H, et al. Promoting wellbeing by changing behaviour: a systematic review and narrative synthesis of the effectiveness of whole secondary school behavioural interventions. Ment Heal Rev J. 2010;15(2):43–53.
15. Durlak J a, Weissberg RP, Dymnicki AB, Taylor RD, Schellinger KB. The impact of enhancing students' social and emotional learning: a meta-analysis of school-based universal interventions. Child Dev [Internet]. 2011 Jan [cited 2011 Feb 15];82(1):405–32. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21291449
16. Sklad M, Diekstra R, De Ritter M, Ben J, Gravesteijn C. Effectiveness of school-based universal social, emotional, and behavioral programs: do they enhance students' development in the area of skills, behavior and adjustment? Psychol … [Internet]. 2012 [cited 2012 Nov 12];49(9):892–909. Available from: http://onlinelibrary.wiley.com/doi/10.1002/pits.21641/full
17. Wilson SJ, Lipsey MW. School-Based Interventions for Aggressive and Disruptive Behavior. Update of a Meta-Analysis. Am J Prev Med. 2007;33(2):S130–43.
18. Slavin RE. Foreword. In: Kelly B, Perkins DF, editors. Handbook of implementation science for psychology in education. Cambridge: Cambridge University Press; 2012. p. xv.
19. National Institute for Health and Care Excellence. Antisocial behaviour and conduct disorders in

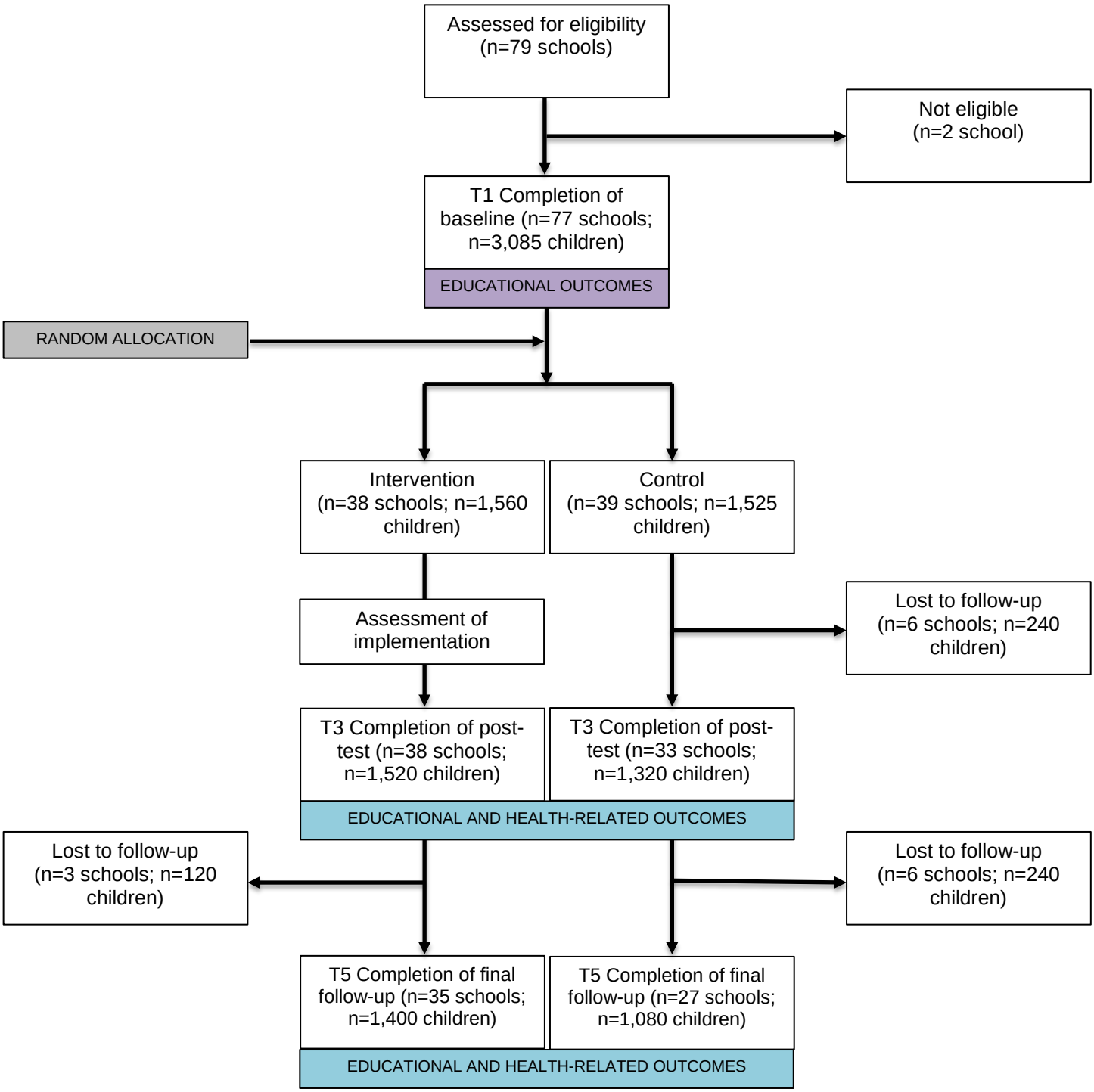
children and young people: recognition, intervention and management. London; 2013.
20. Office for Standards in Education. Below the radar: low-level disruption in the country's classrooms. London; 2014.
21. Department for Education. Mental health and behaviour in schools. London; 2016.
22. Weale S. Schools-based strategy launched to improve children's mental health care. The Guardian. 2014 Nov;
23. Public Health England. The link between pupil health, wellbeing and attainment. London; 2014.
24. Barrish HH, Saunders M, Wolf MM. Good behavior game: effects of individual contingencies for group consequences on disruptive behavior in a classroom. J Appl Behav Anal [Internet]. 1969 Jan [cited 2014 Feb 3];2(2):119–24. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1311049&tool=pmcentrez&rendertype=abstract
25. Kellam SG, Mackenzie ACL, Brown CH, Poduska JM, Wang W, Petras H, et al. The good behavior game and the future of prevention and treatment. Addict Sci Clin Pract [Internet]. 2011 Jul [cited 2014 Feb 3];6(1):73–84. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3188824&tool=pmcentrez&rendertype=abstract
26. Chan G, Foxcroft D, Smurthwaite B, Coombes L, Allen D. Improving Child Behaviour Management: An Evaluation of the Good Behaviour Game in UK Primary Schools. Oxford; 2012.
27. Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. BMJ [Internet]. 2014 Jan 7 [cited 2014 Jul 11];348(mar07_3):g1687. Available from: http://www.bmj.com/content/348/bmj.g1687
28. Tingstrom DH, Sterling-Turner HE, Wilczynski SM. The good behavior game: 1969-2002. Behav Modif [Internet]. 2006 Mar [cited 2014 Feb 3];30(2):225–53. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16464846
29. Elswick S, Casey L. The good behavior game is no longer just an effective intervention for students: An examination of the reciprocal effects on teacher behaviors. Beyond Behav. 2011;21:36–46.
30. Durlak JA, DuPre EP. Implementation matters: a review of research on the influence of implementation on program outcomes and the factors affecting implementation. Am J Community Psychol [Internet]. 2008 Jun [cited 2010 Jul 21];41(3–4):327–50. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18322790
31. US Department of Health and Human Services. Finding the balance: program fidelity and adaptation in substance abuse prevention [Internet]. Retrieved December. 2002 [cited 2014 Sep 16]. Available from: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Finding+the+balance:+program+fidelity+and+adaptation+in+substance+abuse+prevention#1
32. Moore JE, Bumbarger BK, Cooper BR. Examining adaptations of evidence-based programs in natural contexts. J Prim Prev [Internet]. 2013 Jun [cited 2014 Sep 16];34(3):147–61. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23605294
33. Hewitt CE, Kumaravel B, Dumville JC, Torgerson DJ. Assessing the impact of attrition in randomized controlled trials. J Clin Epidemiol [Internet]. 2010 Nov [cited 2014 Nov 26];63(11):1264–70. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20573482
34. Vincent D, Krumpler M. Hodder Group Reading Test manual. London; 2007.
35. Koth CW, Bradshaw CP, Leaf PJ. Teacher Observation of Classroom Adaptation--Checklist: Development and Factor Structure. Meas Eval Couns Dev [Internet]. 2009 Apr 1 [cited 2014 Feb 6];42(1):15–30. Available from: http://mec.sagepub.com/content/42/1/15.refs?patientinform-links=yes&legid=spmec;42/1/15
36. Tschannen-Moran M, Hoy A. Teacher efficacy: Capturing an elusive construct. Teach Teach Educ [Internet]. 2001 [cited 2014 May 25];17:783–805. Available from: http://www.sciencedirect.com/science/article/pii/S0742051X01000361
37. Boyle GJ, Borg MG, Falzon JM, Baglioni AJ. A structural model of the dimensions of teacher stress. Br J Educ Psychol [Internet]. 1995 Mar [cited 2014 May 25];65 (Pt 1):49–67. Available from: http://www.ncbi.nlm.nih.gov/pubmed/7727267

38. Kellam SG, Rebok GW, Ialongo N, Mayer LS. The course and malleability of aggressive behavior from early first grade into middle school: results of a developmental epidemiologically-based preventive trial. J Child Psychol Psychiatry [Internet]. 1994 Feb [cited 2014 May 26];35(2):259–81. Available from: http://www.ncbi.nlm.nih.gov/pubmed/8188798
39. Gneezy U, Leonard KL, List JA. Gender differences in competition: Evidence from a matrilineal and a patriarchal society. Econometrica. 2009;77:1637–64.
40. Lendrum A, Humphrey N. The importance of studying the implementation of school-based interventions. Oxford Rev Educ. 2012;38:635–52.
41. Leflot G, Van Lier P, Onghena P, Colpin H. The role of teacher behavior management in the development of disruptive behaviors: an intervention study with the good behavior game. J Abnorm Child Psychol [Internet]. 2010 Aug [cited 2014 Nov 10];38(6):869–82. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20373016
42. Flower A, McKenna JW, Bunuan RL, Muething CS, Vega R. Effects of the Good Behavior Game on Challenging Behaviors in School Settings. Rev Educ Res [Internet]. 2014 Jun 4 [cited 2014 Jul 30];0034654314536781-. Available from: http://rer.sagepub.com/content/early/2014/06/12/0034654314536781
43. Berkel C, Mauricio AM, Schoenfelder E, Sandler IN. Putting the pieces together: an integrated model of program implementation. Prev Sci [Internet]. 2011 Mar [cited 2014 Sep 16];12(1):23–33. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20890725
44. Bronfenbrenner U. Making human beings human: bioecological perspectives on human development. London: Sage Publications; 2005.
45. Poduska JM, Kellam SG, Wang W, Brown CH, Ialongo NS, Toyinbo P. Impact of the Good Behavior Game, a universal classroom-based behavior intervention, on young adult service use for problems with emotions, behavior, or drugs or alcohol. Drug Alcohol Depend [Internet]. 2008 Jun 1 [cited 2014 Jan 22];95 Suppl 1:S29-44. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2757275&tool=pmcentrez&rendertype=abstract
46. Puffer S, Torgerson DJ, Watson J. Cluster randomized controlled trials. J Eval Clin Pract [Internet]. 2005 Oct [cited 2014 Aug 14];11(5):479–83. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16164589
47. Department for Education and Skills. Primary social and emotional aspects of learning (SEAL): guidance for schools. Nottingham; 2005.
48. Reupert A, Woodcock S. Success and near misses: Pre-service teachers' use, confidence and success in various classroom management strategies. Teach Teach Educ [Internet]. Elsevier Ltd; 2010 Aug [cited 2014 Nov 27];26(6):1261–8. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0742051X10000430
49. Willson VL, Putnam RR. A Meta-analysis of Pretest Sensitization Effects in Experimental Design. Am Educ Res J [Internet]. 1982 Jan 1 [cited 2014 Nov 24];19(2):249–58. Available from: http://aer.sagepub.com/content/19/2/249.short
50. Gorard S. The propagation of errors in experimental data analysis: a comparison of pre- and post-test designs. Int J Res Method Educ [Internet]. Routledge; 2013 Nov [cited 2014 Jul 30];36(4):372–85. Available from: http://dx.doi.org/10.1080/1743727X.2012.741117
51. de Vaus DA. Research design in social research. London: Sage Publications; 2001.
52. Kendal S, Callery P, Keeley P. The feasibility and acceptability of an approach to emotional wellbeing support for high school students. Child Adolesc Ment Health [Internet]. 2011 Nov 10 [cited 2014 May 27];16(4):193–200. Available from: http://doi.wiley.com/10.1111/j.1475-3588.2011.00602.x
53. Bird VJ, Le Boutillier C, Leamy M, Williams J, Bradstreet S, Slade M. Evaluating the feasibility of complex interventions in mental health services: standardised measure and reporting guidelines. Br J Psychiatry [Internet]. 2014 Apr 1 [cited 2014 May 27];204(4):316–21. Available from: http://bjp.rcpsych.org/content/204/4/316.abstract
54. Forman S, Olin S, Hoagwood K, Crowe M. Evidence-based interventions in schools: developers' views of implementation barriers and facilitators. School Ment Health [Internet]. 2009 [cited 2012 Mar 22];1:26–36. Available from: http://www.springerlink.com/index/95gg77kl66n42tpx.pdf
55. Greenberg M, Domitrovich C, Graczyk P, Zins J, Services C for MH. The Study of Implementation in

School-Based Preventive Interventions: Theory, Research, and Practise. Rockville: CMHS; 2005.
56. Terwee CB, Bot SDM, de Boer MR, van der Windt DAWM, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol [Internet]. 2007 Jan [cited 2010 Nov 20];60(1):34–42. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17161752
57. Fink E, Deighton J, Humphrey N, Wolpert M. Assessing the bullying and victimisation experiences of children with special educational needs in mainstream schools: Development and validation of the Bullying Behaviour and Experience Scale. Res Dev Disabil [Internet]. 2015 Jan [cited 2014 Nov 20];36:611–9. Available from: http://www.sciencedirect.com/science/article/pii/S0891422214004624
58. Wolpert M et al. Me and my school: findings from the national evaluation of Targeted Mental Health in Schools. Nottingham; 2011.
59. Stevens K. Developing a descriptive system for a new preference-based measure of health-related quality of life for children. Qual Life Res [Internet]. 2009 Oct [cited 2014 Dec 5];18(8):1105–13. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19693703
60. Ratcliffe J, Stevens K, Flynn T, Brazier J, Sawyer M. An assessment of the construct validity of the CHU9D in the Australian adolescent general population. Qual Life Res [Internet]. 2012 May [cited 2014 Dec 5];21(4):717–25. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21837445
61. Stevens K. Assessing the performance of a new generic measure of health-related quality of life for children and refining it for use in health state valuation. Appl Health Econ Health Policy [Internet]. 2011 May 1 [cited 2014 Dec 5];9(3):157–69. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21506622
62. Stevens K. Valuation of the Child Health Utility 9D Index. Pharmacoeconomics [Internet]. 2012 Aug 1 [cited 2014 Feb 7];30(8):729–47. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22788262
63. Goodman R. The Strengths and Difficulties Questionnaire: A Research Note. J Child Psychol Psychiatry [Internet]. 1997 Jul [cited 2011 Mar 2];38(5):581–6. Available from: http://doi.wiley.com/10.1111/j.1469-7610.1997.tb01545.x
64. Gutman LM, Feinstein L. Children's well-being in primary school: pupil and school effects [Internet]. Childhood A Global Journal Of Child Research. Centre for Research on the Wider Benefits of Learning, Institute of Education, University of London; 2008 [cited 2011 Jan 5]. Available from: http://eprints.ioe.ac.uk/2050/1/Gutman2008Children.pdf
65. Deighton J, Tymms P, Vostanis P, Belsky J, Fonagy P, Brown a., et al. The Development of a School-Based Measure of Child Mental Health. J Psychoeduc Assess [Internet]. 2012 Nov 20 [cited 2013 Jan 17]; Available from: http://jpa.sagepub.com/cgi/doi/10.1177/0734282912465570
66. Gupta SK. Intention-to-treat concept: A review. Perspect Clin Res [Internet]. 2011 Jul [cited 2012 Nov 21];2(3):109–12. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3159210&tool=pmcentrez&rendertype=abstract
67. Cohen J. A power primer. Psychol Bull [Internet]. 1992 [cited 2011 Apr 8];112(1):155–159. Available from: http://pcbfaculty.ou.edu/classfiles/MGT 6973 Seminar in Research Methods/MGT 6973 Res Methods Spr 2006/Week-5 Research Design and Primary Data Collection/Cohen 1992 PB A power primer.pdf
68. Hill C, Bloom H, Black AR, Lipsey MW. Empirical benchmarks for interpreting effect sizes in research. Child Dev … [Internet]. 2008 [cited 2012 Aug 8];2(3):172–7. Available from: http://onlinelibrary.wiley.com/doi/10.1111/j.1750-8606.2008.00061.x/full
69. Peugh JL. A practical guide to multilevel modeling. J Sch Psychol [Internet]. 2010 Feb [cited 2014 Oct 1];48(1):85–112. Available from: http://www.sciencedirect.com/science/article/pii/S0022440509000545

3.8 Appendix

Good Behaviour Game CONSORT diagram



Summary of outcome measure changes from original proposal

	Proposal			**Protocol**			**Rationale**
	Tool	**Items**	**Internal consistency α**	**Tool**	**Items**	**Internal consistency α**	
Mental health	Me and My School scale - MMS (Deighton et al., 2013) Teacher SDQ – Conduct problems Teacher SDQ – Emotional symptoms	18 5 5	.78 (beh) .72 (emo) .74 .78	KIDSCREEN27 – Psychological wellbeing (The KIDSCREEN Group Europe, 2006) Teacher SDQ – Conduct problems Teacher SDQ – Emotional symptoms	7 5 5	.84 .74 .78	The KS is shorter and positively phrased. Allows focus on psychological wellbeing, as difficulties (conduct problems and emotional symptoms) are already assessed with the teacher SDQ.
Resilience	Child and Youth Resilience Measure 12 – CYRM (Liebernberg, Unbar & Leblanc, 2013)	12	.84	KIDSCREEN27 – Social support and peers KIDSCREEN27 – School environment	4 4	.81 .81	The KS is shorter and taps similar constructs. Wording and phrasing of KS more appropriate than CYRM (item references to community and cultural traditions felt to be inaccessible for some children)
Victimisation/ Bullying	Bullying Behaviour and Experience Scale – BBES (Fink, Deighton, Humphrey, & Wolpert, 2015)	13	.86 (vict) .80 (bul)	KIDSCREEN52 – Social acceptance	3	.77	The KS is shorter and taps into victimisation only (this is the construct of interest).
Health-related quality of life	Child Health Utilities 9D – CHU9D (Stevens, 2012)	9		Child Health Utilities 9D – CHU9D (Stevens, 2012)	9		N/A
Total number of items in child self report survey		**52**			**27**		