Drug and alcohol dependence, in press, 2007

Effects of a Universal Classroom Behavior Management Program in First and Second Grades on Young Adult Behavioral, Psychiatric, and Social Outcomes

Sheppard G. Kellam, MD.
American Institutes for Research, 921 E. Fort Avenue, Suite 225, Baltimore, MD 21230
Phone: 410-347-8551, Fax: 410-347-8559, Email: skellam@air.org

C. Hendricks Brown, Ph.D.
Department of Epidemiology and Biostatistics, College of Public Health
University of South Florida, 13201 Bruce B Downs Blvd, Tampa, FL 33612
Phone: 813-974-6672, Email: hbrown@health.usf.edu

Jeanne Poduska, Sc.D.
American Institutes for Research, 921 E. Fort Avenue, Suite 225, Baltimore, MD 21230
Phone: 410-347-8553, Email: jpoduska@air.org

Nicholas Ialongo, Ph.D.
Johns Hopkins University, Bloomberg School of Public Health
624 N. Broadway, 8th Fl., Baltimore, MD 21205
Phone: 410-955-0414, Email: nialongo@jhsph.edu

Wei Wang, Ph.D.
Department of Epidemiology and Biostatistics, College of Public Health
University of South Florida, 13201 Bruce B Downs Blvd, Tampa, FL 33612
Phone: 813-974-6978, Email: wwang@health.usf.edu

Peter Toyinbo, MD.
Department of Epidemiology and Biostatistics, College of Public Health
University of South Florida, 13201 Bruce B Downs Blvd, Tampa, FL 33612
Phone: 813-974-9187, Email: ptoyinbo@health.usf.edu

Hanno Petras, Ph.D.
University of Maryland, Department of Criminology and Criminal Justice
College Park, Maryland
Phone: 301-405-4716, Email: hpetras@crim.umd.edu

Carla Ford, Ph.D.
American Institutes for Research, 921 E. Fort Avenue, Suite 225, Baltimore, MD 21230
Phone: 410-347-8554,Email: cford@air.org

Amy Windham, Ph.D.
American Institutes for Research, 921 E. Fort Avenue, Suite 225, Baltimore, MD 21230
Phone: 410-347-8552, Email: awindham@air.org

Holly C. Wilcox, Ph.D.
Johns Hopkins University, School of Medicine
Hampton House Room 896, Baltimore, MD 21205
Phone: 410-502-0629, Email: hwilcox@jhsph.edu

ABSTRACT

Background: The Good Behavior Game (GBG), a method of classroom behavior management used by teachers, was tested in first and second grade classrooms in 19 Baltimore City Public Schools beginning in the 1985–1986 school year. The intervention was directed at the classroom as a whole to socialize children to the student role and reduce aggressive, disruptive behaviors, confirmed antecedents of later substance abuse and dependence disorders, smoking, and antisocial personality disorder. This article reports on impact to age 19–21.

Methods: In five poor to lower-middle class, mainly African American urban areas, three or four schools were matched and within each set were randomly assigned to one of three conditions: 1) GBG, 2) a curriculum-and-instruction program directed at reading achievement, or 3) the standard program. Balanced assignment of children to classrooms was made and then, within intervention schools, classrooms and teachers were randomly assigned to intervention or control.

Results: By young adulthood significant impact was found among males, particularly those in first grade who were more aggressive, disruptive, in reduced drug and alcohol abuse or dependence disorders, regular smoking, and antisocial personality disorder. These results underline the value of a first-grade universal prevention intervention.

Replication: A replication was implemented with the next cohort of first-grade children with the same teachers during the following school year, but with diminished mentoring and monitoring of teachers. The results showed significant GBG impact for males on illicit drug abuse/dependence disorder. For other outcomes the effects were generally smaller but in the same direction.

Key Words: Good Behavior Game, developmental epidemiology, universal prevention programs, classroom behavior management, drug, alcohol, antisocial behavior, smoking, random effects, additive models.

1. INTRODUCTION

This article reports on the cumulative impact on young adults (ages 19–21) of a universal preventive intervention classroom behavior management program carried out in first and second grades known as the Good Behavior Game (GBG; Barrish et al., 1969). Administered by the teacher classroom-wide, it is directed at socializing children to the role of student and reducing aggressive, disruptive behavior, a well-documented antecedent risk factor for later drug, alcohol, and antisocial behavioral disorders and other problem outcomes. We have conducted three generations of developmental epidemiology–based, randomized field trials involving the GBG in close partnership with the Baltimore City Public School System (BCPSS). In the first generation trial, the GBG was tested by itself and a curriculum/instruction intervention called Mastery Learning (ML; Block and Burns, 1976) was tested separately in a parallel design. The GBG and a curriculum/instruction intervention were combined in the second-generation trial (Ialongo et al., 1999) and further combined with a family/classroom partnership in the ongoing third-generation trial. This report is concerned solely with the outcomes in young adulthood of the first-generation trial of the GBG and is directed at three goals: (1) to report the overall effectiveness and the predicted variation in effectiveness of the GBG within and across outcomes; (2) to report evidence on the theoretically predicted etiological role of early aggressive, disruptive behavior on this profile of outcomes; and (3) to report on the utility of the prevention strategy of aiming an intervention at a single shared early antecedent of a set of later problem outcomes for the purpose of reducing the risk of the entire set of outcomes.

1.1 Developmental Epidemiology and Life Course/Social Field Theory

The Baltimore prevention intervention research and the Chicago/Woodlawn studies that preceded it are grounded in an integration of three scientific paradigms (Kellam and Ensminger, 1980; Kellam and Langevin, 2003; Kellam and Rebok, 1992; Kellam et al., 1999). The first paradigm is *community epidemiology*, which is concerned with understanding the sources of nonrandom distribution of health problems, behaviors, or related factors in a defined community. Community epidemiology provides control of selection bias and, when integrated with the second paradigm, *life course development*, allows the study of variation in developmental antecedents. The third paradigm is the use of a *randomized experiment to test preventive interventions directed at early antecedents of later problem outcomes* to investigate the causal role, malleability, and mediating/moderating effects of risk factors on the course and prevention of behavioral, psychiatric, and social adaptation, and how the impact varies in relationship to a priori prediction. This prevention science strategy has guided our past research and guides the presentation of results and discussion in the present study (Brown and Liao, 1999; Kellam and Langevin, 2003; Kellam and Rebok, 1992; Kellam et al., 1999).

Our research is further grounded in a theoretical view of psychological, behavioral, and social health. *Life course/social field theory* (Kellam et al., 1975) is focused on two dimensions and their interrelationships: the levels of success and failure as defined by an external dimension (i.e. how an individual is viewed by the prevailing society) and an internal dimension (i.e. the psychological, psychiatric well-being of the individual) as well as on the potential reciprocal relationships between the two dimensions. Life course/social field theory is centered on the observation that in each stage of life there are a few main social fields, each with defined social task demands. The adequacy of an individual's responses to these specific social task demands is rated by *natural raters*, such as parents within the family, teachers in the classroom, significant peers in the peer group, or later in the life course, by spouses in the intimate social field or supervisors in the work social field. This process of social task demands and behavioral responses to them is called *social adaptation*; the resulting ratings by the natural rater(s) are termed *social adaptational status* (SAS). In addition to the individual's performance, the natural rater, chance, and the fit of the individual in the social context play roles in an individual's SAS. Aggressive, disruptive behavior, academic problems, and other early antecedents to problem outcomes are viewed, therefore, not as residing merely in the individual but also reflecting the results of social interactions within and across social fields—between child and teacher, classmates/peers, family, and the broader community and societal context.

Life course/social field theory postulates that adapting or maladapting to earlier social task demands in specific social fields leads to later adaptation or maladaptation in the same field as well as in other social fields (Cicchetti and Schneider-Rosen, 1984). The theory also suggests that psychological well-being (PWB), the internal dimension that reflects the psychological status of an individual, may be an antecedent and/or a consequence of social maladaptation, since social maladaptation may be reciprocally related to PWB. For example, failure may make one feel depressed, and/or feeling depressed may make one more likely to fail. Finally, life course/social field theory proposes that improving SAS early in the life course is likely to improve later developmental outcomes.

For several decades, we have studied teachers' social task demands of students in first grade classrooms (Kellam et al., 1975; Werthamer-Larsson et al., 1991). The first social task demand for students is engaging in behavior that is in keeping with classroom rules, to which one maladaptive response is aggressive, disruptive behavior. Other social task demands are that children participate in social interactions with classmates and the teacher and not be too shy or withdrawn, that they pay attention and remain focused, and that they learn the academic subjects. In prior publications we reported that directing interventions at the social adaptational process of social task demands and behavioral responses shows promise for improving both later SAS and PWB. For example, conducting the GBG in first-grade classrooms resulted in decreases in aggression (Dolan et al., 1993) and off-task behavior (Brown, 1993) by the end of

first grade and lower levels of antisocial behavior and tobacco use by middle school (Kellam and Anthony, 1998; Kellam et al., 1994a). In addition, improving achievement (SAS) tends to improve depression (PWB) in girls and to reduce aggression (SAS) in boys (Kellam et al., 1994b; Kellam et al., 1998b) during the course of the first grade. A list of previous publications that have addressed the immediate effects of the GBG by the end of first grade and the intermediate effects by middle school can be found in the online supplementary material[1]. This report, however, focuses on the long-term, young adult outcomes of the GBG intervention as predicted by life course/social field theory and further guided by earlier research on aggressive, disruptive behavior as a risk factor for long-term outcomes.

1.2 Aggressive, Disruptive Behavior

Aggressive, disruptive behavior repeatedly has been shown, as early as the first grade, to be an important maladaptive classroom behavioral antecedent of adolescent and adult illicit drug use, conduct disorders, antisocial personality disorder, criminal behavior, and school failure and dropout (Block et al., 1988; Dishion et al., 1996; Ensminger and Slusarcick, 1992; Ensminger et al., 1983; Farrington and Gunn, 1985; Farrington et al., 1988; Grant, 1991; Hans et al., 1992; Haskins et al., 1983; Kellam et al., 1975; Kellam et al., 1983; Kellam et al., 1991; Kellam et al., 1994a; Kershaw, 1992; Oakes and Lipton, 1990; Patterson et al., 1992; Pekarik et al., 1976; Robins, 1978; Sameroff, 1994; Schwartzman et al., 1985; Shedler and Block, 1991). During the preschool years, this early risk factor is most likely to be found in the interactions of children, their siblings, and their parents in the home setting. In particular, coercive, irritable, and ineffective parenting behaviors have been implicated consistently in the development of conduct problems throughout childhood (McCord, 1988; Patterson et al., 1992; Reid, 1993; Reid and Eddy, 1997) and as a precursor to illicit drug use (Hawkins et al., 1992). McNeil and colleagues (1991), Webster-Stratton (1989, 1998), and Patterson and colleagues (1982) showed that parent-training programs had intervention effects on antisocial behavior, supporting the strategy of directed interventions at these types of antecedents.

When children enter the school system this link between family processes and child outcomes becomes more complex. Preschool-age children who engage in oppositional and aggressive, disruptive behaviors with their parents are at high risk for engaging in aggressive, rejecting behaviors with classmates that not only accelerate the development of antisocial behaviors but also decrease children's ability to profit from positive educational and social opportunities (Reid, 1993). Furthermore, these students may become involved in coercive interactions with their teachers, who often are not trained in classroom behavior management and may inadvertently escalate negative behaviors in children leading to classrooms with higher levels of aggressive, disruptive behavior (Kellam et al., 1998a).

[1] See Publications on GBG Intervention Trial Impact Prior to Young Adulthood in online supplementary material.

These classrooms place children at higher risk of later problem behavior and outcomes, particularly those who themselves are more aggressive, disruptive in first grade (Kellam et al., 1994c; Kellam et al., 1998a). Such classrooms point to the need for interventions that provide teachers with better tools for socializing children into the role of student and for effective classroom behavior management.

1.3 GBG

The GBG was developed by Barrish, Saunders, and Wolf (1969). Before the Baltimore trials, as far as we can determine, no randomized field trials had been conducted of the GBG, but we have found thus far 22 papers and 18 dissertations describing positive results from the use of GBG in pre-post or fairly short term ABAB designed studies with relatively small numbers of children.[2] These observational studies led to the epidemiologically-based randomized trial reported here. In a very recently published review, Tingstrom et al. (2006) infer that the pre-randomized field trials and the Baltimore epidemiologically-based randomized trials support the beneficial effects of GBG as a classroom behavior management tool. Embry (2002) even suggests the GBG as a possible "behavioral vaccine" to prevent subsequent problem behaviors, particularly given the earlier reported results of our randomized trials in Baltimore.

The purpose of the GBG is to create a classroom environment that is conducive to learning for all students. The focus is on the social context of the classroom; the function of GBG is to socialize children into the role of student and to teach them to regulate their own and their classmates' behavior through a process of interdependent team behavior–contingent reinforcement (Tingstrom et al., 2006). The goal of the strategy is to reduce early aggressive, disruptive behavior at the classroom level and at the individual level, a frequently reported antecedent of later problem outcomes. This report is the longest follow-up study of GBG impact yet done, and it provides evidence regarding the effectiveness of GBG.

In the trial reported here, classrooms of first and second graders received the GBG intervention over the course of 2 years. Participating teachers initially received training and assigned children to one of three heterogeneous teams that contained equal numbers of boys and girls, equal numbers of aggressive, disruptive children, and equal numbers of shy, socially isolated children based on baseline measurements of classroom behavior. The teacher posted basic classroom rules of student behavior and teams were rewarded if the team members committed four or fewer infractions of these classroom rules. During a particular game period, all teams were eligible for the reward if they accumulated four or fewer infractions of acceptable student behavior. The GBG was played during those periods of the day when the classroom environment was less structured, such as when the teacher was working with one student or a small

[2] See Lurye I., Mackenzie, A. and Kellam, S.G., History and Evolution of the Good Behavior Game (GBG) in online supplementary material.

group while the rest of the class was instructed to work on assigned tasks independently.

During the first weeks of the intervention, the GBG was played three times each week for a period of 10 minutes. The duration of the game increased approximately 10 minutes per game period every 3 weeks, up to a maximum of 3 hours. Initially, the teacher announced game periods, and the rewards were delivered immediately after the game. Later, the teacher initiated the game periods without announcement, and the rewards were delayed until the end of the school day or the end of the week. Over time, the game was played at different times of the day and during different activities. In this manner, the GBG evolved from a procedure that was highly predictable and visible, with a number of immediate rewards, to a procedure with an unpredictable occurrence and location, with deferred rewards.

We hypothesize that the GBG, directed at the interactive process of the classroom teacher's social task demands and the children's behavioral responses, will be effective overall but that its main impact will be for children who are maladapting to the classroom by aggressive, disruptive behavior, the specific target of GBG. Based on life course/social field theory and earlier results, we further hypothesize that the impact of the GBG will be found among those long-term outcomes predicted by early aggressive, disruptive behavior, generally the more externalizing behaviors. Life course/social field theory also posits that mastery enhances the likelihood of positive outcomes therefore protecting against maladaptive and/or poor psychological outcomes, and so a research question rather than a hypothesis is whether GBG protects children from becoming more at risk over the life course due to the mastery of social adaptation.

2. METHODS

2.1 Epidemiologically Based, Randomized Field Trial Design

Our trial design involved selecting five large urban areas within Baltimore City, matching sets of schools in each area, and randomly assigning which type of intervention would be tested in which elementary schools from these urban areas. We then assigned all of the children entering first grade to classrooms within these schools in a balanced manner, and then randomly assigned the classrooms/teachers to classroom intervention condition. This multilevel design encompassed a total of 19 schools, 41 classrooms, and 1196 children within the 5 urban areas. Control classrooms were assigned both in the same school where the GBG was being used and in other schools. Procedures to ensure balance across types of communities and classroom composition within schools are described below (see Brown et al., in press, for additional details).

The first stage of the design involved selecting the five distinctly different socio-demographic urban areas in Baltimore. Three or four schools were matched in

each of the five urban areas by socioeconomic status (SES), size of school, and ethnicity. These five urban areas, which were selected with the help of city planners and our partners in the BCPSS, varied in SES from very poor to moderate income and in ethnicity, including mostly African American, mixed ethnicities, and mostly white. We then randomly assigned schools within each matched set in each urban area to serve as schools where GBG would be tested (6), schools where ML would be tested independently of GBG (7), or schools where neither of these interventions would be tested (6; external control schools). All of the schools that were used to test GBG or ML had either two or three first-grade classrooms.

The second stage of the design involved assigning individual children to first-grade classrooms within each school so that classrooms were nearly identical before they were assigned to intervention condition. Starting in the summer before the school year began and early in the school year, school administrators assigned all students sequentially using an alphabetized list to the different first-grade classrooms within their school. Classes within each school were checked for balance on kindergarten experience and academic and behavioral performance, and in the few instances (3-4) in which there was imbalance when school began, a small number of children were reassigned. Those children who moved into any of these school catchment areas during the year were assigned sequentially across classrooms, with the provision that the class sizes remained comparable. These procedures produced balanced and equal-size classrooms within the schools.

The third stage of this design was random assignment of classrooms and teachers to intervention condition within each intervention school. Early in the fall of 1985, after the school year started and before the interventions began, we randomly assigned all regular (non-special education) first-grade classrooms along with their teachers to an intervention condition. There were 6 schools across the 5 urban areas in which GBG was tested. A total of 8 GBG classes were randomly selected in these 6 schools. In 4 of these schools there were 2 first-grade classrooms, so 1 classroom in each of these schools was randomly selected to receive GBG and the other served as the control condition. In the 2 remaining GBG schools with 3 first-grade classrooms, 2 were randomly selected to receive GBG and the remaining classroom was assigned to be a control classroom. Therefore, in each of the 6 schools where GBG was being used there was 1 control classroom, referred to as the internal control. Similarly in the 7 schools where ML was being tested, 9 classrooms were randomly selected to serve as ML intervention classrooms and the remaining 7 served as controls for these schools where ML was being tested. In the remaining 6 schools, 11 classrooms served as external controls where no formal intervention was implemented. For the purposes of this article, we do not discuss the 9 ML-assigned classes because their long-term outcomes could be affected by the ML intervention and therefore are not informative in examining the long-term impact of the GBG. However, the ML control classrooms provide an additional control

condition for the GBG testing, as do the classrooms in the external control schools. This design used school as a blocking factor to compare GBG classrooms against internal GBG controls. The use of additional external control classrooms in schools where no intervention was implemented was built into the design because we believed it was possible that the internal control classroom teachers could be influenced by communicating with the GBG classroom teachers in the same school. The disadvantage of these external controls is that school building, family, and community variation would add variability to the GBG versus external control comparison relative to that achievable with comparison to internal control classrooms within the same school. Detailed power calculations strongly favor the use of school as a blocking factor once classrooms within a school are randomly assigned, as was done here (Brown and Liao, 1999). The absence of any evidence of contamination of controls within the same school has led us to examine the impact of the GBG primarily against the internal control classrooms, with the ML and external control conditions allowing important comparisons of consistency of results.

The assignment procedures involved class lists based on the children present after the first 8 to 10 weeks of the first school year when baseline data were obtained before the start of the intervention. A total of 1196 children in the first cohort were available from these lists, and we have used this number to define the original sample that was pursued for follow-up. Our analyses focus on 922 students who were either in GBG classrooms or included in one of the three control groups since the 274 children in ML intervention classrooms were excluded as mentioned earlier. These 922 include the 238 who were assigned to the 8 GBG classrooms and the 169 children assigned to the 6 internal control classrooms in the 6 GBG schools, and the 515 additional children who were in external or ML control classrooms. In our primary analyses we compared the long-term outcomes of the 238 GBG-assigned children to their 169 internal controls, “borrowing strength” in some analyses by using the additional 515 controls to provide supplemental information about classroom and individual levels of variance. In addition, we examined the level of consistency of results by comparing the GBG against all three control groups, separated and combined.

The trial involved 2 years of exposure to the GBG intervention. Children in a GBG class in the first grade (1985–1986) also received GBG during second grade where children’s first-grade classroom assignments remained the same as for the previous year. Teachers of the GBG classrooms received 40 hours of training, most of which occurred at the beginning of the program, followed by supportive mentoring during the course of the school year. A comparable amount of attention was spent with standard program teachers but without a focus on classroom behavior management. These activities were directed at balancing the amount of attention given standard program teachers with that given the GBG teachers.

In the second school year while the first cohort of children were in second grade (1986–1987), another cohort of first graders was assigned in the same balanced fashion to intervention condition and classroom for the first two years of elementary school, creating a second (replication) cohort. For this second cohort, the first-grade teacher remained in the same intervention condition, yet the GBG teachers received little retraining, support or further mentoring and monitoring since we assumed they would continue the intervention with fidelity. More emphasis was placed on training the new second-grade teachers who were continuing with subjects in the first cohort.

2.2 Measure of Aggressive, Disruptive Behavior in First and Second Grades

The Teacher Observation of Classroom Adaptation-Revised (TOCA-R) is a measure of each child's adequacy of performance on the core tasks in the classroom as rated by the teacher. It was developed and used in the Woodlawn studies (Kellam et al., 1975), and after modification, was used as a core periodic assessment instrument for the Baltimore education/prevention trials (Werthamer-Larsson et al., 1991). TOCA-R contains a multi-item scale of each social task demand; each of the roughly 10 items measuring each construct is rated on 6 levels. It involves a 2-hour-long structured interview in a private location in the school, and is administered by a trained member of the staff who initiates the session with the teacher by listening to the teacher's assessment of how the school year is going. After listening and working through trust, the interviewer asks whether the teacher is ready to rate the children. The teacher is then guided in his or her ratings of each child on each item representing the social task demand constructs. The interviewer follows a script precisely, responds in a standardized way to issues the teacher initiates, and records the teacher's ratings of the adequacy of performance of each child in the classroom. The ratings used in analyses in this article are those from the fall of first grade at the time of the first report card.

The SAS construct of central interest here is *Authority Acceptance,* the maladaptive form of which is aggressive, disruptive behavior. Psychometric work includes item–whole correlations among the 10 items for each time of administration, fall and spring in first and second grades and spring each year through seventh grade. The range in alphas was from 0.91 to 0.95. A correlation of 0.67 was found between the Authority Acceptance subscale and peer nominations of "gets into trouble." In terms of predictive validity, the strength of prediction for each young adult outcome is presented in Results. For example, for each unit increase in the TOCA-R Authority Acceptance subscale scores in grade 1, there was about a 50% increase in the likelihood of a diagnosis of antisocial personality disorder at ages 19–21 (odds ratio [OR]=1.49, 95% confidence interval [CI]=1.30–1.71). The 10 TOCA-R items comprising the aggressive, disruptive construct of TOCA-R are *Breaks Rules, Breaks Things,*

Fights, Harms Others, Harms Property, Lies, Stubborn, Teases Classmates, Takes Others Property, and *Yells at Others*.

2.3 Key Outcome Measures Administered at Ages 19–21

A 90-minute-long telephone interview was carried out with each student who participated in the trial, now ages 19–21. Young adults were asked if they would agree to be interviewed and advised that they were not obligated to answer questions if they did not wish to do so, and could terminate the interview at any time. They were offered a $50 participation incentive and were given a T-shirt inscribed with the logo of the prevention program. These procedures were approved by both the Johns Hopkins and the American Institutes for Research institutional review boards. The interviewers were masked to the first-grade intervention condition of the respondents. The questions were organized first by social fields of family of origin, school, work, intimate relationships, family, and peers. This was followed by developmental history and current status, and then by psychiatric diagnosis. The instruments used in analyses reported here were the Composite International Diagnostic Interview-University of Michigan version (CIDI-UM) and the young adult's educational history.

2.3.1 CIDI-UM. CIDI-UM (Kessler et al., 1994), modified to reflect the *Diagnostic and Statistical Manual of Mental Disorders-IV* diagnostic criteria (*DSM-IV;* American Psychiatric Association, 1994), was used to determine the lifetime, past-year, and past-month occurrence of the following: major depressive disorder (MDD), generalized anxiety disorder (GAD), illicit drug abuse/dependence disorder, and alcohol abuse/dependence disorder, antisocial personality disorder (ASPD). We added regular use of tobacco, including number of cigarettes used each day. Diagnoses were derived in accordance with *DSM-IV* criteria, using a computerized scoring algorithm. We present here results on lifetime illicit drug abuse/dependence and alcohol abuse/dependence disorders, MDD, GAD, ASPD and regular use of tobacco.

The CIDI-UM is a fully structured psychiatric interview that specifies the exact wording and sequence of questions and provides a complete set of categories for classifying respondents' replies. The highly structured format is intended to minimize clinical judgment when eliciting diagnostic information and recording responses. To facilitate the accurate recall of lifetime episodes of mental disorders, CIDI-UM includes commitment and motivation probes. The modified CIDI differs from the standard CIDI in placing diagnostic probe questions at the beginning of the interview to minimize response biases associated with fatigue effects. Test-retest and inter-rater reliability studies of the CIDI suggest good to excellent kappa coefficients for most diagnostic sections (Wittchen, 1994). Kessler et al. (1998) reported good to excellent agreement with diagnoses derived by psychiatrists using the Structured Clinical Interview for *DSM-III-R* (SCID; Spitzer et al., 1992).

2.3.2 Young Adult's Educational History. Interviewers asked the young adults to report on the highest level of schooling obtained, the number of repeated grades from K–12, how well they performed overall in school (K–12), whether they were currently in school/training and, if so, how well they were performing, and the nature of the educational program they were currently attending (college, vocational school, etc.).

2.4 Comparison at Baseline and Attrition by Intervention Group and Other Variables

We performed baseline comparisons of individual and classroom-related variables by intervention condition. The results of the first cohort are reported here. The variables used for comparison included the distribution of teacher ratings of behavior on the TOCA-R, student self-reports of psychiatric symptoms, achievement scores, free or reduced-price lunch status by intervention group, classroom size, and gender. These baseline measures were made 10 weeks into first grade, when classroom composition had become fairly stabilized. The intervention began after these baseline measures were taken.

There were no significant differences between the intervention groups on baseline characteristics of teacher ratings of aggressive, disruptive behavior, fall-of-first-grade achievement, or free or reduced-price school lunch when we took into account school as a random factor. For a more stringent comparison of GBG and internal GBG control on baseline characteristics, we repeated these analyses without using random effects of school and classroom. Table 1 shows the overall high comparability of baseline variables between GBG and internal GBG controls, as would be expected from balancing at the classroom level. No baseline comparison approaches significance except for depressive symptoms (p=0.01). We call attention to this difference between GBG and internal controls for depressive symptoms. It was more pronounced among females (p=0.03) compared to males (p=0.08, adjusted for school and classroom effects). While there is a significant difference for depression, there are factors that mitigate its influence on our analyses. First, the significance might be discounted based on the large number of comparisons. Second, depressive symptoms are only correlated 0.02 with aggressive, disruptive behavior, the main risk factor we are targeting, so its impact on our analyses is expected to be small. Third, the effect size difference between GBG and its internal control is only 0.3, thereby diminishing its potential for confounding.

2.4.1 Missing Data on Baseline Teachers' Ratings. All of the children were included who were present at the fall of first grade in the first cohort analyses of the impact of the GBG reported in this article, regardless of their length of exposure to GBG during first and second grades. After examining the simple cross-tabulations of outcomes across intervention conditions, we included baseline teachers' ratings in more elaborate models such as those that include interactions with baseline aggressive, disruptive behavior. As a result, we examined the causes of missing baseline teachers' ratings, whether baseline

ratings differed by intervention group, and whether the baseline ratings themselves related to attrition in the young adult data.

A total of 90% of the children were assessed in the first six weeks of first grade using the TOCA-R (see Table 2, column 3). Twenty percent of those missing baseline teachers' ratings left BCPSS early on had no teachers' ratings throughout elementary and middle school and were overall less likely to be interviewed at follow-up (42%), compared with those who did have teachers' ratings (73% follow-up, $p<0.002$). Aside from these children with little or no data throughout the study, 8% of the sample did not receive baseline teachers' ratings. These include 15 (10%) missing among the internal GBG controls and 7 (3%) missing from the GBG classrooms. An overall test was performed to examine the rate of missing teachers' ratings in GBG and internal GBG controls after adjusting for school. The rate of missing baseline data was slightly higher in the internal GBG control classrooms compared with the GBG classrooms ($p=0.03$ by Mantel-Haenszel test, controlling for school). We found that the vast majority of all of the missing baseline data came from a single school with large classrooms. In this school, 10 of 34 (29%) of children's teachers' ratings were missing in an internal GBG control classroom, and 5 of 35 (15%) were missing in the GBG classroom ($p=0.11$). Both rates are much higher than in all other schools and could be explained by a limited amount of time available to complete the TOCA-R assessments in those classrooms due to the large classroom size.

Baseline teachers' ratings in the fall of first grade were available for 90% of the 922 students (see Table 2). Among the intervention conditions, the external control classrooms had significantly higher rates of missing data ($p<0.001$, unadjusted for schools; see Column 3 of Table 2). There were two reasons for this higher level of missing teachers' ratings in the external control group compared with other groups. First, one of the external control classrooms with 19 students had a change in teacher early in the school year, making baseline teachers' ratings unavailable. Second, nearly all of the remaining missing teachers' ratings occurred in large external control classrooms. Classrooms of 28 children or larger accounted for 86% of the remaining external controls with missing data. By way of comparison with baseline data, at the end of first grade, the rates of completion of teachers' ratings were not significantly different between GBG (78%) and internal GBG control (71%, $p=0.28$) when adjusted for school. Overall, we concluded that the amount of missing data at baseline is modest and the reasons for "missingness" at baseline are likely the result of factors related to classroom size and mobility and not to the interventions themselves.

We note that geographic area was significantly related to the presence of baseline data; again, this can be explained by the change in teacher in one classroom and by larger classrooms with larger amounts of missing baseline data as described above. There is no difference in the presence of baseline data by free or reduced-price school lunch status or gender. Table 2 shows that

females were more likely to be interviewed at follow-up compared with males ($p<0.001$), and rates of follow-up differed by urban area in first grade ($p\leq0.001$). Intervention status, however, was unrelated to attrition ($p=0.81$). Likewise, baseline scores on teachers' ratings, self-reports of psychiatric symptoms, free lunch status, and achievement scores were unrelated to interview status at young adult follow-up. We also compared the rates of missingness by intervention in either the baseline data or the young adult interview; having missing data in either of these would exclude the case from the two-level mixture models described below. Controlling for school, there was a non-significant difference between GBG and internal GBG control on this overall missingness index ($p=0.15$ by Mantel-Haenszel test).

2.5 Statistical Methods

We have provided three types of intent-to-treat analyses because there are no published and available statistical methods that can handle all of the complexities that we would like to include simultaneously in our models. Each provide legitimate statistical tests of impact and address multilevel data, but each do so in different ways by emphasizing different aspects of our theory guided intervention and providing different levels of statistical power to examine key elements of our model. The first method involves two-level modeling of child- and classroom-level effects. We attempted to fit a three-level mixed effects model adding urban area as a fixed factor, with individual, classroom, and school as the three-level random effects; however, existing commercial software could not give us reliable results. The computationally feasible two-level models, with individual- and classroom-level effects, provide the most direct test of the theory of what impact GBG would have on youth with different risks of aggressive, disruptive behavior. We rely on this model to examine both linear and nonlinear variation in impact based on individual-level risk. We have included classroom-level random effects and, where necessary, classroom-level predictors in these analyses because the group-based randomized trial involved random assignment at the classroom level (Murray, 1998). Thus, these two-level linear and additive logistic regression models with random intercepts provide adjustments for potential dependence of observations at the classroom level and for baseline-by-treatment interactions. In addition, we have carried out multiple imputations to account for missing data on baseline aggressive, disruptive behavior, depressive symptoms, and distal outcomes (Schafer, 1997).

The second and third analytical methods—one based on Mantel-Haenszel statistics and the other on a paired *t* test for log ORs—were used to condition on the school to account for the blocking of this factor in the design. These two analyses for binary outcomes provide overall effects of the impact of GBG, but because of the sparse data, they cannot take into account individual-level risk, which is central to the theory driving the intervention. Thus, these two methods provide a legitimate overall test of impact with small-sample corrections based on the number of schools, but they fail to take into account variation in individual-

level risk and have low statistical power to detect intervention effects when there is variation in impact as a function of individual-level risk.

2.5.1 Analyses Using Two-Level Modeling with Individual-Level Risk and Random Classroom Effects to Assess Impact of GBG. We used large sample tests on two-level modeling because few if any exact tests are available for dichotomous outcomes fit with linear logistic and nonlinear random effects modeling. These tests are generally appropriate when the number of subjects is large, the number of classes is moderate, and the magnitude of the intra-class correlations is small, as we have found empirically to be the case in these analyses. Models that involve logistic regression analyses are fit using S plus version 7 (Insightful, Seattle, WA). These models include main and interactive effects involving the baseline levels of aggressive, disruptive behavior, intervention condition, and gender if either linear interactions or nonlinear interactions are indicated (Brown, 1993). Those with nonlinear interactions are fit in R version 2.1.1 (R Development Core Team), with generalized additive models (GAM) using the logistic link function (Dominici et al., 2002; Hastie and Tibshirani, 1990). Further details regarding the analytic methods are provided in the following section.

We present the results in detail for the first cohort because it was the original trial in which teachers of GBG classrooms received the intended appropriate training, mentoring, and monitoring. Summaries are also provided of the impact on the second cohort, the replication trial where teachers received less mentoring and monitoring and no retraining. The identical analytical strategy was used for the second cohort analyses as was done for the first. Complete descriptions, tables, and figures for the second cohort are available in the supplementary material on this journal's website.[3]

3. RESULTS

Formal analyses of GBG intervention against internal GBG controls were conducted as a series of parallel analyses, one for each young adult outcome. For each outcome, we begin by reporting simple cross-tabulations of outcome by intervention status to show the overall magnitude of the effect. These cross-tabulations compare GBG against the internal controls as well as all controls. We also compare the group's rates for each outcome for those who score above 3.5 (roughly top 12% for boys, 3% for girls) on aggressive, disruptive behavior in first grade because we are hypothesizing differential impact for this higher-risk group. By using the same cut-off score for both boys and girls we are able to compare gender effects on the same scale. (The nonlinear models and figures that follow provide a more detailed examination of variation in impact). These simple summaries are followed by more sophisticated analyses that take into account child baseline data and classroom variation. To arrive at our inferences about the GBG effect on each separate young adult outcome, we used a two-step strategy: model selection followed by formal hypothesis testing. The first

[3] See online supplementary material. http://www.sciencedirect.com/science/journal/03768716.

step, model selection, requires screening through more than 12 candidate fixed-effect models to determine whether, for example, males and females are more appropriately examined separately or together, whether baseline interactions are required, and whether nonlinear or linear logistic regression models should be used. The best-fitting model for testing each separate outcome is selected without regard to intervention effect parameters so that testing for intervention impact can be done independently of the decisions required for obtaining best-fitting models. The second step, hypothesis testing, begins with the best-fitting candidate model, then includes random effects for all classrooms, and formal tests for the impact of the intervention using Wald tests and likelihood ratio tests.

Regarding specifics of the first step of selecting a candidate model, all model comparisons were based on likelihood ratio tests of nested models. For each outcome variable, we computed a comprehensive series of analyses that compared males' and females' responses to the GBG condition as well as to each of the control conditions. All of these models included the main effects of gender and log-transformed baseline levels of teachers' ratings of aggressive, disruptive behavior in the fall of first grade; both of these effects were frequently found to be significant for the outcomes included here. Models with both linear and nonlinear effects for baseline aggressive, disruptive behavior were then examined. Likelihood ratio tests were used to compare the fit of linear logistic models with comparable GAM models whose degrees of freedom for nonlinear covariates were set to 3 for each continuous covariate or interaction term. We then examined models that included interactive effects of gender and baseline aggressive, disruptive behavior, and main effects as well as linear and nonlinear interactive effects of intervention condition and baseline, and linear and nonlinear three-way effects of gender, intervention condition, and baseline. We included in these models separate contrasts between the three control conditions, thus allowing for variations that took place across the different controls. To identify best-fitting candidate models, we used likelihood ratio tests under independence assumptions to first check for linear and nonlinear interaction terms.

Sequentially, the above steps examined the necessity of including gender-by-baseline interactions or three-way interactions, then determined whether nonlinear models were required, then determined whether baseline-by-intervention analysis was required. The final step examined the impact of the intervention. The *p* value cutoffs for testing all interactions and nonlinear effects were set at 0.10, and those involving main effects (and GBG effects) were set at 0.05.

Once we identified a best-fitting candidate model, we added classroom random effects. We also repeated these analyses using school rather than classroom as a random factor, but found that classroom-level analyses produced a better fit. If the variance of the classroom-level random effect was large, as we noted in the initial analyses of high school graduation, then we investigated whether inclusion of family or community poverty measures provided additional explanatory value.

If the final model was a generalized linear model, then we compared the fit using a generalized linear mixed-effects model (GLMM; Breslow and Clayton, 1993; Wolfinger and O'Connell, 1993) with classroom random effects. If the final model included nonlinear predictors, that is, a generalized additive model, we then compared the fitting result with the corresponding generalized additive mixed-effects model (GAMM; Wang, 1998; Wood, 2004) with classroom-level random effects.

Testing was accomplished by computing the GBG impact effect against internal control classrooms. Formal tests of intervention impact relied on both Wald-type tests of intervention effect relative to its standard error and likelihood ratio tests, in which the contrast between GBG and internal control condition was dropped. Wald-type tests of intervention impact were based on critical values from a *t*-distribution, with degrees of freedom determined by the number of degrees of freedom at the classroom level rather than the number of degrees of freedom at the child level. For testing involving GAMM, maximum likelihoods were not available in R, so they were computed directly through numerical integration using Mathematica 5.2 (Wolfram Research, Champaign, IL). To test whether variation in classroom effects was negligible, we compared the likelihood ratio of the mixed model to a model without these random effects and compared the result to a mixture of χ^2 distributions (Stram and Lee, 1994). Even if the random effects were not significant, we performed all tests with this random effect in the model. For the first cohort, we did not find that the introduction of classroom effects changed our conclusions about the impact of the GBG except for graduation from high school. Classroom variations did affect some analyses in the second cohort. To verify whether conclusions about impact were influenced by missing data on baseline aggressive, disruptive behavior, baseline depressive symptoms, and distal outcomes, we used multiple imputation procedures with the imputation model including these same measures repeated over time (Schafer, 1997). Formal comparisons were based on Wald-type tests that accounted for within and between variations across imputations and adjusted for the random effects of classroom.

We present results of GAM models for all plots, even if they did not significantly improve over the standard logistic models. These GAM plots provide smooth nonlinear fits that provide useful visual information about the impact of the GBG that complements the formal statistical tests. For models for which there were no interactions involving gender, we collapsed the fits for males and females because the overall fits resulted in similar-shape curves. For models with significant gender-by-intervention or baseline effects, we present separate GAM plots for each gender. The plots compare the GBG directly with its own internal GBG control as well as with the two other control conditions in different schools. Overall, these external controls and ML controls provide additional confirmation of the impact of the GBG on many of the outcomes.

3.1 Assessing Impact of GBG Using Within-School Variation and Small-Sample Testing

Small-sample testing was used because the number of GBG classrooms in our study (8) was small as were the number of schools where GBG classrooms were present (6). These small-sample tests address the fact that intervention assignment was at the classroom level and schools served as a blocking factor. They ignore individual-level data, however. The large-sample and small-sample tests have complementary strengths and in general result in similar conclusions when baseline-by-intervention interactions are not present.

We note, first, that the two-level mixed logistic regression models described in the previous section do not incorporate all of the hierarchical structure in the data because blocking exists within some of the schools, and schools are nested within geographic regions of Baltimore City. Second, the traditional test statistics that are used are large-sample test statistics; these have excellent properties when the number of independent units is large, but because our classroom-based intervention trial involves a total of 8 GBG classrooms and 6 internal GBG control classrooms, all within 6 schools, these tests may have higher type I error rates than the nominal 0.05 level we used. To examine these two issues, we provide two conditional tests that perform better in small samples.

The first is a test of no association between intervention condition (GBG versus internal GBG control) and young adult outcome using a continuity-adjusted Mantel-Haenszel test (Agresti, 1990) that specifically accounts for the blocking of intervention classrooms within schools. By stratifying on school in comparing the GBG versus internal GBG control children, the Mantel-Haenszel test assumes a common OR for intervention condition by young adult outcome and conditional independence of children within schools. Such a model is appropriate if there is small variation in these ORs because of classroom variation, which we have empirically determined to be the case.

Finally, we provide an analysis that specifically addresses the small number of GBG and internal GBG control classrooms. This analysis is based on a paired *t* test involving the six school-level log ORs (corrected for small sample size by the addition of ½ to each cell); that is, each school's log OR estimates the difference in log odds for GBG versus internal GBG controls. This last test of no association, which is based on 5 degrees of freedom, is likely to be the least powerful of our tests, not only because it is a small-sample test but also because it ignores information that would clearly increase power—it does not "borrow strength" from the variation found in the other 18 control classrooms nor does it use any of the child-level data in predicting outcome. For the young adult outcomes that show no significant baseline-by-treatment interaction, we carry out this *t* test on log ORs based on the entire sample. For the young adult outcomes that show baseline-by-treatment interactions, we have typically not calculated log ORs restricted to those children scoring above a certain cutoff point. This is

because the choice of a cutoff point is rather arbitrary, and the data become too sparse to be able to use the normal approximation for the log OR.

We note that to date there are no statistical programs that handle the full four-level nesting of this design in logistic regression. We have confidence, however, that the addition of these classroom-level random effects and the models that condition specifically on the level of the school provide reliable inferences of intervention effect in such four-level data. Brown (1993) has examined four-level data on off-task behavior from this same experiment and concluded that adjustment for classroom-level variation alone is sufficient to obtain good-quality inferential statements.

3.2 Effects of the GBG on Young Adult Outcomes at Ages 19–21

Our examinations of the effects of the GBG on young adult outcomes are based on exposure to the GBG in first and second grades and self-report data collected during young adulthood. These comparisons of rates for GBG versus control conditions were made for males and females combined and also separately by gender, where the gender differences are meaningful. We have previously found that the GBG generally had a greater impact on those who were more aggressive, disruptive in first grade compared with those who exhibited lower levels. Therefore, we also present rate comparisons for GBG versus controls among those who had been the most highly aggressive, disruptive first graders.

Using simple cross-tabulation summaries, we begin by reporting all rate differences that are statistically significant without adjusting for classroom-level effects. Thus, the overall rate comparisons by intervention condition are appropriate for overall comparisons, but once the classroom effects are controlled for the *p*-values for intervention significance may be reduced. These overall comparisons begin with analyses that combine males and females, and then compare them separately by gender and for those who had the highest levels of aggressive, disruptive behavior in first grade. Non-significant trends—those with significance levels >0.05—are also noted. We follow this with the more extensive large-sample and small-sample analyses described above.

3.3 Lifetime Illicit Drug Abuse/Dependence

Using simple cross-tabulations, we found among the first cohort that young adults who were in GBG first-grade classrooms had lower levels of illicit drug abuse/dependence disorders compared with those in the control classrooms (12% for GBG vs. 21% for internal controls, p=0.04, and 19% for all controls, p=0.03, unadjusted for child baseline and classroom effects). This reduction was significant for males; 19% of GBG males reported illicit drug abuse/dependence compared with 38% of internal male controls (p=0.01) and 30% of all control males (p=0.05, unadjusted for child baseline and classroom effects). GBG did not reduce illicit drug dependence or abuse in females (8% vs. 7% for internal controls p=0.93, 11% for all female controls, p=0.40, unadjusted for child

baseline and classroom effects). We found much greater effects of the GBG for males among those who were rated as being highly aggressive, disruptive by teachers in first grade. For these highly aggressive, disruptive males, the rate of illicit drug dependence or abuse was 29% in the GBG groups, 83% in the internal controls (p=0.02), and 68% in the combined control group (p=0.02, unadjusted for child baseline and classroom effects). These initial findings are supported by more complex analyses described below.

3.3.1 Impact of GBG on Illicit Drug Abuse/Dependence Examined With Individual-Level Risk Factors and Classroom Variation. For lifetime illicit drug abuse/dependence, the best-fitting model included main effects of gender and baseline aggressive, disruptive behavior with no indications of nonlinearity. For example, there was no indication of an interactive effect of GBG by baseline aggressive, disruptive behavior (p=0.22, unadjusted for classroom effects). Therefore, we used logistic regression models with classroom random effects to test for the main effects of GBG versus controls, along with intervention by baseline effects. We also found significant interactions involving gender and baseline, and we present separate analyses for males and females.

In the GLMM analysis presented in Table 3 for males, which adjusts for classroom variation, the reduction in lifetime illicit drug abuse/dependence for GBG compared with internal GBG controls was large and significant (log OR of 0.999, p=0.035). This corresponds to an approximately 2.7 times greater risk of abuse or dependence in the internal GBG controls compared with GBG. There was no indication of differences in rates among the three control groups. When we compared the most general GAMM models across all intervention groups, GBG rates for illicit drug abuse/dependence were lower than any of the control groups across all levels of baseline (Figures 1 and 2). The effect of the GBG on males is much stronger than it is on females as shown by the non-significant effect for GBG versus internal GBG control in Table 4 and the distinct differences between GBG and controls across baseline for males (Figure 2) but not for females (Figure 3). A GLMM was used to test for non-zero random classroom variation. For both males and females, these classroom effects were minimal. The model fit using numerical integration methods (e.g., adaptive Gaussian quadrature method) resulted in estimated parameter values that were almost identical to the result from the generalized linear model run without classroom effects. To insure that this important GBG impact was not influenced by the baseline imbalance in depressive symptoms we included baseline depressive symptoms as a covariate. We used multiple imputations to handle missing data. The GBG effect became more pronounced (log OR=-1.216, SE=0.421), yielding a more significant finding (p=0.008).

The next results of testing for intervention impact in males on lifetime illicit drug abuse/dependence explicitly take into account the blocking factor of school. The Mantel-Haenszel continuity-corrected test that conditioned on school found a significant lower probability of illicit drug abuse or dependence for males exposed

to GBG compared with internal GBG controls (χ^2=5.16, 1 *df*, *p*=0.023). The corresponding small-sample test of the value of the log OR for males, accounting for school as a blocking factor, also indicated significance (*t*=-2.72, 5 *df*, *p*=0.042) with a 95% CI for the OR of 0.11–0.94. Thus, all three analyses indicate a significant reduction in males' illicit drug abuse/dependence for GBG compared with controls.

3.3.2 Impact on Illicit Drug Abuse/Dependence in the Second Cohort. In the second cohort, there was also a reduction in illicit drug abuse/dependence for GBG males (19%) compared to internal controls (32%, *p*=0.10, not controlling for classroom effects). The overall benefit of the GBG compared to low aggressive, disruptive internal controls was significant at *p*=0.016 but not compared to all controls (24%, *p*=0.41). Using the same model selection procedure as for the first cohort, the covariate terms in the best-fitting model for Cohort 2 were close to those in the first cohort analyses. Furthermore, the beneficial impact of the GBG on illicit drug abuse/dependence was found for males using GAMM. Yet unlike Cohort 1, the impact in Cohort 2 was most prominent among the lower aggressive, disruptive youth (*p*=0.016) and diminished somewhat as baseline aggressive, disruptive behavior increased (*p*=0.05). Overall, the nonlinear smoothers show that for three quarters of the sample, GBG showed a consistent gain in preventing this disorder compared to matched controls, and for the remaining quarter the fits showed very similar risks. Small-sample testing of the impact of the GBG showed that this effect was in the same direction but not significant (χ^2=2.25 on 1 *df*, *p*=0.133 for Mantel-Haenszel test; *t*=-1.64 on 5 *df*, *p*=0.16 for paired *t*-test).

3.4 Lifetime Alcohol Abuse/Dependence

We return now to Cohort 1. Using simple cross-tabulations, we found a reduction in lifetime alcohol abuse/dependence in the GBG group compared with controls (13% for GBG versus 20% for internal controls, *p*=0.08, and 29% for all controls, *p*=0.03, unadjusted for baseline or classroom effects). The overall effects of GBG appeared to be similar for both males and females.

3.4.1 Impact of GBG on Lifetime Alcohol Abuse/Dependence Examined With Individual-Level Risk Factors and Classroom Variation. For lifetime alcohol abuse/dependence, the best-fitting model included significant main effects for gender, baseline aggressive, disruptive behavior, and intervention group. Baseline aggressive, disruptive behavior was related to alcohol dependence or abuse, and it showed a moderate departure from linearity, so we included an additive term for baseline aggressive, disruptive behavior in modeling lifetime alcohol abuse or dependence. We tested GBG versus internal GBG control with a single contrast in this model, and found an overall significant reduction in the log odds of an alcohol diagnosis (log OR=-0.70, standard error [SE]=0.35, *p*=0.05; see Table 5). This effect size implies a 50% reduction in the OR of a lifetime alcohol abuse/dependence diagnosis. A GAM plot to examine this risk

reduction as a function of baseline aggressive, disruptive behavior is provided in Figure 4. We note that the variation in impact appears most pronounced throughout the middle level of baseline risk; GAM models, however, show no significant interactive effect between intervention condition and baseline.

When we examined GBG versus the separate GAM fits for the three control groups, we found that the risk was consistently lower for GBG than all three controls for all values along the baseline of aggressive, disruptive behavior (Figure 4). We also fit a generalized random effects model based on the best-fitting model that included gender, baseline aggressive, disruptive behavior and intervention group, where baseline aggressive, disruptive behavior appeared as an additive term. We did not find that incorporating classroom effect altered the significance of these effects.

The pattern of the impact of GBG on alcohol abuse or dependence by gender was similar to that shown for illicit drug abuse or dependence. The effects for GBG for females appeared somewhat smaller than that for males, but this gender by intervention difference was non-significant.

3.4.2 Impact of GBG on Lifetime Alcohol Abuse/Dependence Examined With Within-School Variation and Small-Sample Testing. When we combined genders, the Mantel-Haenszel continuity-corrected test that conditioned on school did not find a significant association between GBG and a lower probability of alcohol abuse/dependence (χ^2=1.63, 1 *df*, *p*=0.20). The corresponding small-sample test of the value of the log OR accounting for school as a blocking factor showed near-significant results (*t*=-2.18, 5 *df*, *p*=0.081), with a 95% CI for the OR of 0.39–1.08. The Mantel-Haenszel test on males alone provided significant results (χ^2=4.66, 1 *df*, *p*=0.03). The corresponding small-sample test for males of the value of the log OR accounting for school as a blocking factor yielded (*t*=-3.69, 5 *df*, *p*=0.014) with a 95% CI for the OR of 0.20–0.75.

3.4.3 Impact on Alcohol Abuse/Dependence in the Second Cohort. In the second cohort, the mixed model logistic regression analyses indicated no significant impact on alcohol abuse/dependence, in contrast to the finding in the first cohort.

3.5 Lifetime Regular Smoking

Using simple cross-tabulations, we found that the rate of regular smoking, defined as ≥10 cigarettes per day at the time of the young adult interview, was lower among those in GBG classrooms (6%) than it was among those in control classrooms (10% for internal controls, *p*=0.15, and 14% for all controls, *p*=0.002, unadjusted for baseline or classroom effects). This effect was higher among GBG males—6% were regular smokers compared with 19% of male internal controls (*p*=0.03) and 20% of all control males (*p*=0.004, unadjusted for baseline or classroom effects)—than it was for females—5% were regular smokers

compared to 3% of internal female controls (p=0.55) and 9% of all female controls (p=0.20; unadjusted for baseline or classroom effects). These reductions in regular smoking were highly apparent among males who began first grade with high aggressive, disruptive teachers' ratings; for the high aggressive, disruptive males, none were regular smokers in the GBG groups (0%) compared with 40% of internal controls (p=0.008) and 25% in the all control groups (p=0.03, unadjusted for classroom effects).

3.5.1 Impact of GBG on Regular Smoking Examined With Individual-Level Risk Factors and Classroom Variation. More complex analyses reveal a significant linear interaction of GBG with baseline ratings of aggressive, disruptive behavior (log OR=-8.81, SE=3.45, p=0.01), as well as significant three-way interactions between baseline, intervention, and gender (log OR=3.62, SE=1.76, p=0.04). These analyses indicate that there is a significant difference in the impact of GBG involving gender and baseline aggressive, disruptive behavior. These findings support results from our most general GAM models displayed in Figure 5, in which smoking for males in the GBG group is compared with that of males in the other treatment groups. Therefore, separate analyses for each gender with correction for classroom effects were performed. For males who smoked ≥10 cigarettes/day, the best-fitting model included main effects for baseline aggressive, disruptive behavior plus interactive effects of baseline by intervention condition. Because there is no significant departure from linearity, we report the results of the logistic regression mixed effects model of intervention impact. Classroom random effects were found to be relatively small. The result listed in Table 6 has been controlled for classroom effects. There is a significant interactive effect of GBG on baseline aggressive, disruptive behavior compared with its internal control (log OR=-6.92, SE=2.09, p=0.001). A likelihood ratio test confirmed a highly significant effect of GBG when compared with the internal GBG control (p=0.003). Overall, GBG is associated with a reduction in the probability of males smoking ≥10 cigarettes/day, and the effect of GBG is greater for boys with higher levels of aggressive, disruptive behavior in first grade compared with lower levels where the rates of regular smoking are similar.

For females, however, the best-fitting model is the logistic model with no interaction of baseline aggressive, disruptive behavior and intervention. With a correction for classroom effects (see Table 7), there is no significant main effect of GBG compared with its internal control group (log OR=0.68, SE=1.13, p=0.55). GAM plots also reflect this minimal difference in the smoking pattern between GBG females and their control counterparts.

3.5.2 Impact of GBG on Regular Smoking Examined With Within-School Variation and Small-Sample Testing. Our small-sample test of the log ORs that accounts for the school blocking factor did not confirm significant effects of GBG for combined males and females when the GBG group was compared with the internal controls (t=-1.07, 5 df, p=0.33). This test does show a marginally significant decrease in risk for regular smoking for GBG males over internal

control males (mean log OR=-1.19, *t*=-2.43, 5 *df*, *p*=0.06; OR=0.30 [95% CI 0.09–1.07]). The effect of the GBG is also close to significance for high aggressive, disruptive males and females combined (mean log OR=-1.40, *t*=-2.86, 4 *df*, *p*=0.06; OR=0.25 [0.06–1.06]), although the data are sparse. The effects of the GBG on high-risk males are significant based on a Mantel-Haenszel continuity-corrected test (χ^2=4.22, 1 *df*, *p*=0.04). Both tests found no significant effect of GBG in females regardless of their ratings for baseline aggressive, disruptive behavior.

3.5.3 Impact on Smoking in the Second Cohort. In the second cohort, both GBG males and females showed 30% to 40% lower rates of regular smoking compared with their counterparts; this reduction was non-significant but consistent across baseline levels of aggressive, disruptive behavior as well as control conditions. In the final mixed logistic regression models, these effects were non-significant for both males (*p*=0.26) and females (*p*=0.52).

3.6 Lifetime ASPD

Overall rates of ASPD were lower for youth in the GBG groups (17%) than they were for internal controls (25%, *p*=0.07) and all controls (25%, *p*=0.03, unadjusted for classroom- and child-level effects). The GBG also reduced ASPD among males who were rated as highly aggressive, disruptive in first grade compared with similar high aggressive, disruptive males in the control groups (38% for GBG aggressive, disruptive males versus 80% for internal controls, *p*=0.10, and 70% for all aggressive, disruptive male controls, *p*=0.05, also unadjusted for classroom- and child-level effects).

3.6.1 Impact of GBG on ASPD Examined With Individual-Level Risk Factors and Classroom Variation. For ASPD, the best-fitting candidate model included significant main effects for gender, baseline aggressive, disruptive behavior, as well as baseline-by-gender and baseline-by-intervention group interactions. We therefore used the GAMM function provided by R package MGCV to fit a GAMM based on the final model, which included gender, baseline aggressive, disruptive behavior, intervention group, baseline-by-gender interaction and baseline-by-GBG interaction. In this analysis we used a total of 3 degrees of freedom for additive terms. Table 8 shows all of the terms in this model, including the main effects of gender and baseline level of aggressive, disruptive behavior, and nonlinear main effects and interactive effects.

In analyzing ASPD, we found that males overall, as well as those with higher levels of aggressive, disruptive behavior at baseline, predicted higher levels of ASPD as expected. Also, there were two-way interactions between baseline aggressive, disruptive behavior and gender and between the baseline and GBG that showed marked departure from linearity, so we included additive terms for these two interaction effects in modeling ASPD.

The overall comparison of GBG versus internal GBG controls on ASPD showed a significant effect that varied by baseline level of aggressive, disruptive behavior. The combination of a main effect for this contrast between GBG and internal GBG control (Table 8, row 2) and its nonlinear interaction with baseline (row 7) correspond to a total χ^2 of10.05 with 3 *df* and a *p*-value of 0.018, demonstrating that baseline by GBG effects are strong. They are shown in Figure 6 in GAM plots. We found that the random effect of classroom was so small that its incorporation in the model did not alter the significance of these effects (Table 8, row 10). There were no differences between the three control groups (Table 8, rows 3, 4, 8, and 9).

Figure 6 shows overall that the GBG has more impact on ASPD for students with lower ratings of aggressive, disruptive behavior but has markedly more impact on those at the upper end with high aggressive, disruptive behavior. There is a small, non-significant crossing of the curves in the middle of this scale. Figure 6 presents a comparison of GBG against the separate control groups. The pattern of benefit for GBG with low and high levels of aggressive, disruptive behavior at baseline is consistent across all three control groups.

3.6.2 Impact of GBG on ASPD Examined With Within-School Variation and Small-Sample Testing. Our results of testing for intervention impact on ASPD explicitly took into account the blocking factor of school. When we combined females and males, the Mantel-Haenszel continuity-corrected test that conditioned on school did not find a significant association between the GBG and a lower probability of ASPD (χ^2=2.30, 1 *df*, *p*=0.13). The corresponding small-sample test of the value of the log OR, accounting for school as a blocking factor, was also not significant (*t*=-1.77, 5 *df*, *p*=0.14). For males, the Mantel-Haenszel test did not reach significance (χ^2=1.59, 1 *df*, *p*=0.21). The corresponding *t* test of the log ORs for each school gave *t*=-1.71, 5 *df, p*=0.15. The effect on females was weaker (*t*=0.48, 5 *df*, *p*=0.62). Our GAM analyses have isolated the intervention benefit to high-risk as well as low-risk males. The small-sample tests limited to these groups are too sparse and provide a less-reliable assessment than our GAM modeling did.

3.6.3 Impact on ASPD in the Second Cohort. In Cohort 2, the impact of GBG on ASPD was not significant. Although much less in magnitude, GAM plots showed a similar pattern of benefit among individuals at the highest risk level.

3.7 High School Graduation

Using simple cross-tabulations, we found that overall high school graduation rates were somewhat higher for youth exposed to the GBG (72%) than for internal controls (64%, *p*=0.16) and all controls (65%, *p*=0.11, not correcting for classroom variation). These differences were stronger among males (68% for GBG males, 54% for internal controls (*p*=0.09) and 57% for all controls (*p*=0.08) than they were for females (74% for GBG, 72% for internal controls, *p*=0.79, and

72% for all controls, p=0.63). The differences in graduation rates were most striking for males in GBG first-grade classrooms who had been highly aggressive, disruptive (75%) compared with the highly aggressive, disruptive males in control classrooms (20% in internal controls, p=0.03, and 40% in all controls, p=0.04, unadjusted for classroom and baseline effects). There were no significant differences in high school graduation rates for females regardless of baseline risk.

3.7.1 Impact of GBG on High School Graduation Examined With Individual-Level Risk Factors and Classroom Variation. For high school graduation, the best-fitting model included significant main effects for gender, baseline aggressive, disruptive behavior, as well as baseline-by-intervention interaction, and a borderline significant gender-by-baseline interaction (p=0.10), so we examined the impact of GBG on high school graduation separately by gender and combined with covariates (see Table 9). For both males alone and for males and females combined, a nonlinear GAM analysis shows significant improvement over its linear version.

Unlike the analyses for all of the other outcomes in this article, we found that classroom variation in rates of high school graduation were significantly different from zero. Because high school completion rates vary dramatically by social class and ethnicity in urban areas, we included several additional individual-level and contextual-level predictors as possible explanatory variables. These included free or reduced-price lunch as an index of family poverty, race/ethnicity, classroom proportion of free or reduced-price lunch, and urban area. There were two urban areas that had significantly lower levels of high school graduation than the other three areas (57% versus 85%; p<0.01). Two of the three urban areas had high levels of poverty; every classroom from these two areas had a high level of free or reduced-price lunch, while every classroom from the other three areas had a low level of free or reduced-price lunch. Inclusion of either urban area or classroom poverty reduced the classroom random effect to non-significant (p=0.50, SD=0.007; the point estimate without this SES predictor is 0.852, p<0.001).

Whether or not these individual-level and contextual effects were included in our model with random classroom effects, the overall improvement introduced by the GBG is not significant (χ^2=2.23, 3 *df*, p=0.53 for males and females combined, and similarly non-significant for males with individual and classroom poverty level included). We found no significant effect of the GBG either as a main effect or in interaction with individual-level gender, aggressive, disruptive behavior, or poverty, or with classroom-level poverty or aggressive, disruptive behavior. The interaction of the GBG with baseline for males, shown in row 6 of Table 9, fails to identify any significant impact; however, there was a consistent pattern of higher high school graduation rates for the GBG compared with controls among those with high initial levels of aggressive, disruptive behavior in first grade (Figure 7). This pattern appeared consistently across most of the geographic regions and

within high aggressive, disruptive as well as low aggressive, disruptive behavior groups. We found, however, that such baseline-by-intervention effects were non-significant throughout all of our comparisons.

3.7.2 Impact of GBG on High School Graduation Examined With Within-School Variation and Small-Sample Testing. We also conducted tests for intervention impact on high school graduation that explicitly took into account the blocking factor of school. When we combined females and males, the Mantel-Haenszel continuity-corrected test that conditioned on school did not find a significant association between GBG and a higher probability of high school graduation (χ^2=1.05, 1 *df*, *p*=0.30). The corresponding small-sample test of the value of the log OR, accounting for school as a blocking factor, also was not significant (*t*=-1.63, 5 *df*, *p*=0.16) with a 95% CI for the OR of 0.70–5.01. For males, the Mantel-Haenszel test did not reach significance (χ^2=2.43, 1 *df*, *p*=0.12; *N*=134 in 6 schools). The corresponding *t* test of the log ORs for each school was also non-significance (*t*=1.800, 5 *df*, *p*=0.13).

3.7.3 Impact on Graduation in the Second Cohort For the second cohort, there was no significant difference between GBG and internal GBG controls on graduation rates, although the results favored GBG (64% verses 56% for males, *p*=0.39).

3.8 Lifetime Generalized Anxiety Disorder (GAD)

The overall rates of GAD in this sample were small and did not significantly differ by intervention condition based on simple cross-tabulations (2% for GBG, 3% for internal controls, *p*=0.37, and 2% for all controls, *p*=0.78, not controlling for classroom effects). The differences for males were in a different direction (0% for GBG, 4% for internal control, *p*=0.10, and 2% for all controls, *p*=0.19) than they were for females (3% for GBG, 2% for internal controls, *p*=0.99, and 2% for all controls, *p*=0.52, not controlling for classroom effects).

3.8.1 Impact of GBG on GAD Examined With Individual-Level Risk Factors and Classroom Variation. For lifetime GAD, the best-fitting model included significant main effects for gender, baseline aggressive, disruptive behavior, and gender-by-intervention interaction. A nonlinear GAM model does not show significant improvement over its linear version; therefore, we included only linear terms in this logit model. Also, increased aggressive, disruptive behavior was associated with higher levels of anxiety later on, and females were more likely than males to have GAD.

Because of the interaction between gender and intervention, we performed separate analyses for males and females. Neither of the groups showed a significant effect of the GBG. There was a trend toward lower levels of GAD for high aggressive, disruptive males who were exposed to GBG as compared with controls.

3.8.2 Impact of GBG on GAD Examined With Within-School Variation and Small-Sample Testing. For small-sample testing, the paired *t* test for males involving log ORs reached a *p* value of 0.06 (*t*=-2.44, 5 *df*). The Mantel-Haenszel test was non-significant (*p*=0.40). Together, these tests are suggestive of effects of the GBG on males, but possibly because of the low overall base rate, they are not definitive.

3.8.3 Impact on GAD in the Second Cohort. In Cohort 2, the rates of GAD were quite similar to that in the first cohort and did not significantly differ across intervention condition (1.8% for GBG and 1.6% for all controls, *p*=0.92).

3.9 Lifetime Major Depressive Disorder (MDD)

The unadjusted rates of lifetime MDD were slightly lower overall for the GBG group (10%) compared with internal controls (15%, *p*=0.27), and compared with all controls (12%, *p*=0.52, not controlling for classroom effects). For males' MDD rates, the differences were slightly larger than those for females; male rates were 9% for GBG versus 14% for internal controls (*p*=0.35) and 10% for all male controls (*p*=0.78) whereas the corresponding rates for females were 12% for GBG, 15% for internal controls (*p*=0.49) and 14% for all controls (*p*=0.52, not controlling for classroom effects). The fact that male MDD rates are only modestly lower than females in this Baltimore sample is somewhat surprising. Research utilizing the CIDI-UM to measure MDD found that among 19-21 year olds of a largely Hispanic population in South Florida, females had approximately twice the prevalence of MDD among males (Turner and Gil, 2002), rather than the 50% higher rates reported here. The gender difference in MDD rates is well reported in adults and has been shown to emerge in early adolescence (Weller et al., 2006, Wade et al., 2002, Lewinsohn et al., 1998); however relatively few gender comparisons exist for urban, African American populations similar to this study in Baltimore. The Epidemiological Catchment Area (ECA) studies of the early 1990s reported adult rates of MDD to be similar between genders in Baltimore, while in the other five catchment areas around the country there were differences between genders (Romanoski et al., 1992), pointing to the possibility that ethnic and/or geographic differences may influence this gender difference.

3.9.1 Impact of GBG on Lifetime MDD Examined With Individual-Level Risk Factors and Classroom Variation. For predicting MDD in males and females combined, the linear logistic is the best-fitting model and includes main effects for gender, baseline aggressive, disruptive behavior, and intervention group, plus two-way interactions between gender and aggressive, disruptive behavior at baseline, and three-way interactions. Therefore, separate analyses were performed for both males and females. For males, the best-fitting model is the model containing fixed effects of baseline aggressive, disruptive behavior and intervention and incorporating random effects of the classroom. Baseline aggressive, disruptive behavior is a significant predictor of lifetime MDD in males (*p*=0.01). No significant effects of GBG compared with the internal control group were found (*p*=0.30).

The best model for females is an additive model that includes the intervention condition and baseline aggressive, disruptive behavior in a nonlinear fashion. There are also no significant effects for GBG compared with the internal control group (p=0.23).

3.9.2 Impact on MDD in the Second Cohort. For the second cohort, the effect of GBG on MDD for males was marginally significant (3.6% for GBG versus 12% for the internal GBG control, p=0.05 without adjusting for classroom effects). There was no evidence for and effect on females (12% for GBG versus 7.8% for their respective controls, p=0.41). Analyses taking into account classroom variation did not detect a significant intervention effect.

4. DISCUSSION

The effects of GBG on young adult outcomes suggest that directing a universal intervention at the first grade teacher and classroom to improve socializing children into the role of student and classroom behavior management has both immediate and long-term benefits particularly for males and even more so for those at higher levels of aggressive, disruptive, disruptive behavior early in first grade. GBG was directed at improving early mastery of appropriate student behavior, thereby improving the developmental trajectories and preventing a set of problem outcomes, mainly those that could be thought of as externalizing outcomes. Table 10 summarizes our significant findings, with fitted proportions calculated in each intervention condition based on individual- and classroom-level analytical models. We emphasize these particular two-level model results because they incorporate our theory of individual-level maladaptive aggressive, disruptive, disruptive behavior embedded in the classroom social adaptational process involving teacher and classmates, thus providing the clearest test of our theory.

Life course/social field theory postulates that aggressive, disruptive, disruptive maladaptive behavior in the social field of the classroom will lead to long term poor outcomes in later social fields, and that directing an intervention, such as the GBG, at the social adaptational classroom process will ameliorate such outcomes. This central concept of the theory suggests a hypothetical mechanism for how this two-year intervention in first grade yielded such long-term outcomes in young adulthood. Children who are successfully taught how to master the social task demands of the first grade classroom may carry with them to higher age and grade levels the experience of mastery (“I can do it”). It is also helpful to look back at the immediate earlier effects of the GBG to assess potential mechanisms for these long-term effects. Besides the obvious advantages of improved behavior in first grade and during elementary school, the GBG also reduced off-task behavior, thus allowing teachers more time/opportunity to teach (Brown, 1993) and the beneficial consequences of mastery of academic subjects such as reading skills required for further mastery later on. The results reported here are consistent with life course/social field

theory as regards the importance of early mastery of social task demands in the classroom in promoting later successful social adaptation not only in school but also in other main social fields.

The present trial had a limited ability to carry out more detailed mediational models to examine specific mechanisms leading to improved child outcomes. We plan to report on the role of early responses to GBG in mediating long-term outcomes, but other analyses will be required to more fully test these and other mediational processes. For example, the current third-generation trial in Baltimore incorporates independent observation of children's behavior, child outcomes, and the quality and fidelity of implementation of teachers' classroom behavioral management.

The intent-to-treat analyses reported here by design examine two single measurement times separated by 14 years. Thus, the malleability of the relationship between first grade and young adulthood is summarized in these analyses without a developmental examination. That is, these analyses are notably blind to detecting impact through changes in growth trajectories and the known interrelationships of SAS and PWB over time, especially those involving the co-occurrence of substance use and mental disorders. Other analyses reported previously (Muthén et al., 2002; Wang et al., 2005) address developmental questions through different forms of growth mixture analyses. These developmental models have demonstrated that the GBG had a greater impact among initially high aggressive, disruptive males through middle school. This impact on the initial high aggressive, disruptive males also holds up for high school graduation when examined with general growth mixture models, although the intent-to-treat analyses reported here did not reach significance.

There are clear, consistent findings in the results for Cohort 1 across these young adult outcomes, although findings differ somewhat in the details. We found that GBG had a significant and substantial impact among all of the externalizing behaviors. The impact of GBG was overall highest among higher aggressive, disruptive, disruptive first-grade males compared with less or moderately aggressive, disruptive, disruptive males. In addition, the effects of GBG on all categories of substance abuse and dependence were consistently strong, whereas the effects on psychological disorders in these intent-to-treat analyses were consistently non-significant. In general, risk did not increase with baseline aggressive, disruptive behavior for females, and for all outcomes, the effects appeared stronger for males than for females.

Life course/social field theory does not explain gender differences in development or in impact of the GBG. These GBG outcome differences suggest that early developmental processes that are salient for males may be different than those for females. The differences may be critical in understanding the gender differences in developmental and in intervention outcomes, and in directing preventive interventions at those that are more germane to females.

Early aggressive, disruptive, disruptive behavior may not have the same developmental salience for females as for males. We reported in 1983 on gender differences in the Woodlawn longitudinal study that among girls, more developmental continuity from first grade through adolescence was found in internalizing symptoms such as depression and anxiety as compared with boys. Among boys, early aggressive, disruptive behavior was more developmentally predictive of later teenage aggression and drug use from first grade to ages 16–17 (Kellam et al., 1983; Ensminger et al., 1983). We have noted some gender differences in these data; for example, the predictability of later external behavior problems as a function of first-grade aggressive, disruptive behavior is far stronger for males than females. Such differences in predictability do not seem to be explained by the generally lower level of aggressive, disruptive behavior in girls compared with boys, although the problem may lie in our aggressive, disruptive, disruptive measures being more fitting for males than females.

There were some other differences in impact across outcomes as well. The two outcomes most closely related to illegal behavior—illicit drug abuse/dependence and ASPD—demonstrated the highest impact among those with the highest aggressive, disruptive behavior in first grade. The effects on males' high school graduation resembled that for ASPD, but the effect was not significant. One distinction concerning high school graduation compared with other outcomes was that community SES played an important predictive role. It may be that the effects of the GBG in schools with high poverty are difficult to detect because of other major influences. It should also be kept in mind that in contrast to the externalizing outcomes, high school graduation may depend more on the course of aggressive, disruptive, disruptive behavior, not merely the rating of it in the fall of the first grade. In any case, more research is needed to explain the variation at the community level combined with the impact of GBG.

We found null effects of GBG on anxiety and depressive disorders. Similar to findings in our previous studies, we note that at a population level, early antecedents of depressive and anxious symptoms are distinct from those of externalizing behaviors (Kellam et al., 1983). We note that the pattern of impact on GAD, although non-significant, is intriguing. Because the statistical power in these models depends on the prevalence of the outcome (as well as the strength of association between baseline aggressive, disruptive behavior and the outcome), the low frequency of GAD may partly explain why we failed to find effects for this outcome. By using symptom- and/or syndrome-level measures, we may well find significant impacts of intervention on internalizing psychiatric disorders. Furthermore, GBG may have an impact on depression or anxiety disorders through a reduction in co-occurrence and co-morbidity, a topic we will examine more directly later. Because these other analyses will follow, we have not used any corrections for multiple comparisons or individual-level analyses here. It is likely that the statistical analyses reported here are not independent across outcomes.

For the second cohort, we concluded that there was a somewhat dissimilar pattern of replication of beneficial GBG findings for illicit drug abuse/dependence. In Cohort 2 the impact was the strongest at the lower levels of aggressive, disruptive, disruptive behavior while in Cohort 1 the effect was apparent across all levels and strongest at the higher levels of aggressive, disruptive, disruptive behavior. Furthermore, there was a similar pattern of GBG benefit for regular smoking, but it was non-significant. There is a suggestion of beneficial impact on males' MDD rates. The second cohort analyses for other outcomes yielded null findings. There are two major hypotheses for these diminished findings in the second cohort. First, GBG showed a much reduced effect on short-term aggressive, disruptive, disruptive behavior in the second cohort compared with the first, and our earlier finding of the impact of GBG on adolescent smoking appeared not to be mediated through aggressive, disruptive behavior (Kellam and Anthony, 1998). We hypothesize that GBG was implemented in the second cohort with less precision than it was in the first cohort because we did not have in place sufficient mentoring and monitoring procedures, resulting in two types of consequences: a diminished impact and a shift in impact. If this reasoning is correct, it would explain the less significant impact and/or a shift in which children were benefited from more highly aggressive, disruptive to less aggressive, disruptive in the second cohort. Another important area for further study is the weaker relationship in the second cohort between aggressive, disruptive behavior in first grade and adult behavioral outcomes. There is a hypothetical statistical explanation for these diminished findings in the second cohort. Classroom variability and heterogeneity across the three control conditions were markedly higher in the second cohort compared to the first so that statistical power was reduced.

Although we did not measure the quality of GBG implementation in this first generation of trials, we strongly hypothesize that the GBG must be carried out with precision including continuing mentoring and monitoring. In the second generation of trials in the early 1990s where we did measure implementation, the results revealed marked reduction in impact when that intervention (combined classroom behavior management and enhanced curriculum/instruction) was done with less precision (Ialongo et al., 1999). This problem of low sustainability without continuing mentoring and monitoring is being reported by other investigators and is the new research frontier as we move through the phases of prevention research from efficacy through effectiveness into the problem of sustainability of new practices and ultimately into system-wide fidelity as programs are disseminated (Elliott and Mihalic, 2004; Hallfors and Godette, 2002). Olds and colleagues (2003) have reported success using an ongoing monitoring system that provides information on fidelity to mentors as well as to the researchers who remain in close partnership with agencies carrying out the Nurse-Family Partnership program. In the current, third-generation GBG trial in Baltimore, we are testing a multilevel mentoring and monitoring structure while we move the program toward dissemination as the data warrant.

From a prevention science and policy perspective there are important lessons if these results are replicated in places other than Baltimore. First, the overall strategy of directing a universal intervention at a shared antecedent of a set of later problem outcomes appears to have been successful. This success provides grounds for real optimism that universal prevention strategies such as the GBG can be done early and economically and can address a set of outcomes, not just one at a time. The importance of this finding cannot be overemphasized. It suggests that each problem outcome may not require a separate early preventive intervention, but that a set of outcomes can respond to an intervention aimed at a single shared antecedent risk factor.

Second, the GBG generally had its strongest effect on the highest-risk youth. We have reported elsewhere that such children, particularly males in highly aggressive, disruptive, disrupted classrooms, are the children most at risk for continuing and later problem behavior (Kellam et al., 1998a). Providing the teacher with tools for socializing children into the role of student and managing his or her classroom appears to reduce the high risk for these early aggressive, disruptive, disruptive children, and demonstrates the utility of this universal intervention for maintaining such children in the mainstream classroom and helping them to develop successfully.

Third, the results point to new areas for training new teachers and in-service training for more experienced ones. Teachers are not as often trained in classroom behavior management as these results strongly suggest they should be. Our data indicate that up to half of the teachers are not prepared to manage their first-grade classrooms effectively, and effective tools such as the GBG can and must be provided to teachers (Kellam et al., 1998a).

Fourth, the results underline the vital importance of the first-grade classroom as a social field where developmental trajectories are displayed and further shaped beyond the earlier social fields of family or preschool settings. This is in no way meant to deny the importance of earlier interactions between parents and children and preschool teacher and students. Social adaptation in first grade left to itself without intervention is strongly predictive of later social adaptation, but by first grade there is still considerable malleability—room for improvement—particularly among higher-risk males. Universal interventions in the first-grade classroom can be decisive in setting the direction for life course in school and beyond. The strength and clarity of the impact of this precisely directed early intervention emphasizes the validity of early universal interventions in this critical social field. Finding the most salient early antecedents for girls will allow extending the testing of this strategy.

Fifth, continual monitoring and mentoring at multiple levels of a school district are likely requirements of moving to high levels of sustained effectiveness in interventions in classrooms. The No Child Left Behind Act of 2001 (PL 107-110) emphasizes information systems that can play a vital role in monitoring and

mentoring. The current high level of attention being paid to measuring school building levels of achievement should also focus on classrooms and individual children at the micro level, as well as to city- and statewide information systems at the more macro levels. A multilevel monitoring system can provide a framework for understanding what problems exist and where. Randomized field trials can provide answers to what works, for which children, under what circumstances. Such trials should be conducted within a multistage epidemiological framework, however. Statewide information systems allow for demographic epidemiology of where the nonrandom distribution of problems is occurring. A second-stage representative sample can provide a sampling frame for causal modeling to examine potential early risk factors that are then potential targets for preventive interventions. The third stage, the randomized field trial, can then be conducted on a representative third-stage sample directed at the early antecedent.

In the end, there is still a great need for replicating the GBG and other universal prevention programs in other school districts with similar and different social, ethnic, and economic characteristics. Even as we produce evidence of effective programs, we face new challenges: how to sustain teacher practices over consecutive cohorts of children and how to go to scale within and across school districts. Such questions need to be addressed at the same time that we are replicating trials of the GBG and other programs in different contexts. We have already learned in this first-generation trial that merely training teachers for 1 year is not sufficient support to sustain new practices. The third-generation trial under way in Baltimore is testing a multilevel mentoring and monitoring structure, and similar and different designs for dissemination are needed. How to test these emerging designs is a challenge to the public health prevention and education research fields.

Ultimately, prevention research in the public education and public health fields needs to be integrated. The risk factors for outcomes in each field overlap and are in many cases the same. Both require basic partnerships between scientists and school districts. This particular study is the product of such a partnership. To develop a design and carry out randomization at multiple levels as described here required a continuing partnership with the BCPSS, just as was required in the early Woodlawn studies beginning in 1964 (Kellam, 2000). The studies reported here and elsewhere from the Baltimore partnership have been based on strong mutual self-interests of the BCPSS, the families and their children, and the researchers. We have functioned over the 21 years of our work together as a research and development arm of BCPSS, while carrying out basic and applied prevention research. The ongoing third-generation trial in Baltimore is part of the BCPSS Master Plan, and it represents the kind of real-world mutual benefit that can occur with such a partnership.

ACKNOWLEDGMENTS

Since 1984, with the strong support of Ms. Alice Pinderhughes, the much-respected superintendent during the early years of the study, the partnership between our research team and the Baltimore City Public School System (BCPSS) has made possible three generations of randomized field trials. This partnership has been a firm foundation for much research serving the mutual interests of the BCPSS and the prevention and education research fields. Dr. Leonard Wheeler, Area Superintendent, the school principals, and the teachers played vital roles. During the follow-up into young adulthood, Dr. Patricia Welch, Chair of the Board of School Commissioners played a major supportive role. The families were incredibly supportive and provided written consent for those parts of the studies requiring extra psychological and other measures beyond the usual academic ones. The young adults who began as the first- and second-grade students were cooperative and collaborated fully in the follow-up measurement of outcomes.

Dr. Jaylan Turkkan did a superb job training the teachers in the GBG and supporting them during the course of the first year of the trial. Dr. Joseph Brady was an early advisor and supporter and instrumental in recruiting Dr. Turkkan into the important role she played. While Director of the National Institute on Drug Abuse (NIDA), Dr. Charles R. Schuster made the bold decision to augment the National Institute of Mental Health (NIMH) prevention research center grant with the necessary support to carry out the trial. Dr. Lawrence Dolan was overall intervention chief during the trial and made major contributions to its rigor and fidelity. Dr. James Anthony made major contributions to the design and early implementation of the trial. Dr. Lisa Ulmer led the early assessment team and made important contributions to the framing of the research. We thank Dr. Jon Baron for valuable comments on several versions of this manuscript and thank members of the Prevention Science and Methodology Group for their helpful discussions that led to improvements in the paper. The authors are much indebted to Amelia Mackenzie who provided extensive and very insightful editorial and research assistance throughout the long process of checking and double checking data, references, and inferences, as well as for improving the readability of the text.

During the last 21 years this research has been supported by NIMH Grants R01 MH 42968, P50 MH 38725, R01 MH 40859, and T32 MH018834, with supplements from NIDA for each of the cited research grants.

REFERENCES

Agresti, A., 1990. Categorical Data Analysis. Wiley, New York.

American Psychiatric Association, 1994. Diagnostic and Statistical Manual of Mental Disorders, IV ed. Washington, DC.

Barrish, H., Saunders, M., Wolf, M., 1969. Good behavior game: effects of individual contingencies for group consequences on disruptive behavior in a classroom. J. Appl. Behav. Anal. 2, 119–124.

Block, J., Block, J.H., Keyes, S., 1988. Longitudinally foretelling drug usage in adolescence: early childhood personality and environmental precursors. Child Dev. 59, 336–355.

Block , J., Burns, R., 1976. Mastery Learning. In: Shulman, L. (Ed.), Review of Research in Education, Vol. 4. F.E. Peacock, Itasca, IL, pp. 3-49.

Breslow, N.E., Clayton, D.G., 1993. Approximate inference in generalized mixed models. J. Am. Stat. Assoc. 88, 9–25.

Brown, C.H., 1993. Statistical methods for preventive trials in mental health. Stat. Med. 12, 289–300.

Brown, C.H., Kellam, S. G., Ialongo, N., Poduska, J., Ford, C., (in press). Prevention of aggression behavior through middle school using a first grade classroom-based intervention. Proc. Annu. Meet. Am. Psychopathol. Assoc.

Brown, C.H., Liao, J., 1999. Principles for designing randomized preventive trials in mental health: an emerging developmental epidemiology paradigm. Am. J. Community Psychol. 27, 673–710.

Cicchetti, D., Schneider-Rosen, K., 1984. Toward a transactional model of childhood depression. In: Cicchetti, D., Schneider-Rosen, K. (Eds.), Childhood Depression: A Developmental Perspective. Jossey-Bass, San Francisco, pp. 5–28.

Dolan, L.J., Kellam, S.G., Brown, C.H., Werthamer-Larsson, L., Rebok, G.W., Mayer, L.S., Laudolff, J., Turkkan, J., Ford, C., Wheeler, L., 1993. The short-term impact of two classroom-based preventive interventions on aggressive and shy behaviors and poor achievement. J. Appl. Dev. Psychol. 14, 317-345.

Dishion, T.J., Spracklen, K.M., Andrews, D.W., Patterson, G.R., 1996. Deviancy training in male adolescent friendships. Behav. Ther. 27, 373–390.

Dominici, F., McDermott, A., Zeger, S.L., Samet, J.M., 2002. On the use of generalized additive models in time-series studies of air pollution and health. Am. J. Epidemiol. 156, 193–203.

Embry, D.D., 2002. The good behavior game: a best practice candidate as a universal behavioral vaccine. Clin. Child Fam. Psychol. Rev. 5, 273-297.

Elliott, D.S., Mihalic, S., 2004. Issues in disseminating and replicating effective prevention programs. Prev. Sci. 5, 47–53.

Ensminger, M.E., Kellam, S.G., Rubin, B.R., 1983. School and family origins of delinquency: comparisons by sex. In: Van Dusen, K.T., Mednick, S.A. (Eds.), Prospective Studies of Crime and Delinquency. Kluwer-Nijhoff Publishing, Boston, pp. 73–97.

Ensminger, M.E., Slusarcick, A.L., 1992. Paths to high school graduation or dropout: a longitudinal study of first grade cohort. Sociol. Educ. 65, 95–113.
Farrington, D.P., Gallagher, B., Moorley, L., St. Ledger, R.J., West, D.J., 1988. Are there any successful men from criminogenic backgrounds? Psychiatry. 51, 116–130.
Farrington, D.P., Gunn, J., eds., 1985. Aggression and Dangerousness. John Wiley & Sons, New York.
Grant, T.M., 1991. The legal and psychological implications of tracking in education. Law Psychol Rev. 15, 299–312.
Hallfors, D., Godette, D., 2002. Will the “Principles of Effectiveness” improve prevention practice? Early findings from a diffusion study. Health Educ. Res. 17, 461–470.
Hans, S.L., Marcus, J., Henson, L., Auerbach, J.G., Mirsky, A.F., 1992. Interpersonal behavior of children at risk for schizophrenia. Psychiatry. 55, 314–335.
Haskins, R., Walden, T., Ramey, C.T., 1983. Teacher and student behavior in high- and low-ability groups. J. Educ. Psychol. 75, 865–867.
Hastie, T., Tibshirani, R., 1990. Generalized Additive Models. Chapman & Hall, London.
Hawkins, J.D., Catalano, R.F., Miller, J.Y., 1992. Risk and protective factors for alcohol and other drug problems in adolescence and early adulthood: implications for substance abuse prevention. Psychol. Bull. 112, 64-105.
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. Am. J. Community Psychol. 27, 599–641.
Kellam, S.G., 2000. Community and institutional partnerships for school violence prevention. In: Preventing School Violence: Plenary Papers of the 1999 Conference on Criminal Justice Research and Evaluation: Enhancing Policy and Practice through Research, Vol 2. National Institute of Justice, Washington, DC, pp. 1–21.
Kellam, S.G., Anthony, J.C., 1998. Targeting early antecedents to prevent tobacco smoking: findings from an epidemiologically based randomized field trial. Am. J. Public Health 88, 1490–1495.
Kellam, S.G., Branch, J.D., Agrawal, K.C., Ensminger, M.E., 1975. Mental Health and Going to School: The Woodlawn Program of Assessment, Early Intervention, and Evaluation. University of Chicago Press, Chicago.
Kellam, S.G., Brown, C.H., Rubin, B.R., Ensminger, M.E., 1983. Paths leading to teenage psychiatric symptoms and substance abuse: developmental epidemiological studies in Woodlawn. In Guze, S.B., Earls, F.J., Barrett, J.E. (Eds.), Childhood Psychopathology and Development. New York: Raven Press, New York, pp. 17-51.
Kellam, S.G., Ensminger, M.E., 1980. Theory and method in child psychiatric epidemiology. In: Earls, F. (Ed.), International Monograph Series in

Psychosocial Epidemiology, Vol. 1: Studying Children Epidemiologically. Neale Watson Academic Publishers, New York, pp. 145–180.
Kellam, S.G., Koretz, D., Moscicki, E.K., 1999. Core elements of developmental epidemiologically based prevention research. Am. J. Community Psychol. 27, 463–482.
Kellam, S.G., Langevin, D.J., 2003. A framework for understanding “evidence” in prevention research and programs. Prev. Sci. 4, 137–153.
Kellam, S.G., Ling, X., Merisca, R., Brown, C.H., Ialongo, N., 1998a. The effect of the level of aggression in the first grade classroom on the course and malleability of aggressive behavior into middle school. Dev. Psychopathol. 10, 165-185. [See also Kellam, S.G., Ling, X., Merisca, R., Brown, C.H., Ialongo, N., 2000. The effect of the level of aggression in the first grade classroom on the course and malleability of aggressive behavior into middle school: Results of a developmental epidemiology-based prevention trial: Erratum. Dev. Psychopathol. 12, 107.]
Kellam, S.G., Mayer, L.S., Rebok, G.W., Hawkins, W.E., 1998b. Effects of improving achievement on aggressive behavior and of improving aggressive behavior on achievement through two preventive interventions: An investigation of causal paths. In: Dohrenwend, B. (Ed.), Adversity, Stress, and Psychopathology. Oxford University Press, London, pp. 486–505.
Kellam, S.G., Rebok, G.W., 1992. Building developmental and etiological theory through epidemiologically based preventive intervention trials. In: McCord, J., Tremblay, R.E. (Eds.), Preventing Antisocial Behavior: Interventions from Birth through Adolescence. The Guilford Press, New York, pp. 162–195.
Kellam, S.G., Rebok, G.W., Ialongo, N., Mayer, L.S., 1994a. 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 35, 259–282.
Kellam, S.G., Rebok, G.W., Mayer, L. S., Ialongo, N., Kalodner, C.R., 1994b. Depressive symptoms over first grade and their response to a developmental epidemiologically based preventive trial aimed at improving achievement. Dev. Psychopathol. 6, 463–481.
Kellam, S.G., Rebok, G.W., Wilson, R., Mayer, L.S., 1994c. The social field of the classroom: context for the developmental epidemiological study of aggressive behavior. In: Silbereisen, R.K., Todt, E. (Eds.), Adolescence in Context: The Interplay of Family, School, Peers, and Work in Adjustment. Springer-Verlag, New York, pp. 390–408.
Kellam, S.G., Werthamer-Larsson, L., Dolan, L.J., Brown, C.H., Mayer, L.S., Rebok, G.W., Anthony, J.C., Laudolff, J., Edelsohn, G., Wheeler, L., 1991. Developmental epidemiologically-based preventive trials: baseline modeling of early target behaviors and depressive symptoms. Am. J. Community Psychol. 19, 563-584.
Kershaw, T., 1992. The effects of educational tracking on the social mobility of African Americans. J. Black Stud. 23, 152–169.

Kessler, R., McGonagle, K., Zhao, S., Nelson, C.B., Hughes, M., Eshleman, S., Wittchen, H., Kendler, K., 1994. Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States. Arch. Gen. Psychiatry 51, 8-19.

Kessler, R.C., Wittchen, H.U., Abelson, J.M., McGonagle, K.A., Schwarz, N., Kendler, K.S., Knauper, B., Zhao, S., 1998. Methodological studies of the Composite International Diagnostic Interview (CIDI) in the U.S. National Comorbidity Survey. Int. J. Methods Psychiatr. Res. 7, 33–55.

Lewinsohn, P.M., Rohde, P., Seeley, J.R., 1998. Major depressive disorder in older adolescents: prevalence, risk factors, and clinical implications. Clin. Psychol. Rev. 18, 765-794.

Lurye, I., Mackenzie, A., Kellam, S.G., in preparation. History and evolution of the good behavior game (GBG).

McCord, J., 1988. Parental behavior in the cycle of aggression. Psychiatry. 51, 14–23.

McNeil, C.B., Eyberg, S.M., Eisenstadt, T.H., Newcomb, K., Funderburk, B.W., 1991. Parent-child interaction therapy with behavior problem children: generalization of treatment effects to the school setting. J. Clin. Child Psychol. 20, 140-151.

Murray, D.M., 1998. Design and Analysis of Group-Randomized Trials. Oxford University Press, New York.

Muthén, B.O., Brown, C.H., Masyn, K., Jo, B., Khoo, S.T., Yang, C.C., Wang, C.P., Kellam, S.G., Carlin, J.B., 2002. General growth mixture modeling for randomized preventive interventions. Biostatistics 3, 459–475.

Oakes, J., Lipton, M., 1990. Tracking and ability grouping: a structural barrier to access and achievement. In: Goodlad, J., Keating, P. (Eds.), Access to Knowledge: An Agenda for Our Nation's Schools. College Entrance Examination Board, New York, pp. 187–204.

Olds, D., Hill, P., O'Brien, R., Racine, D., Moritz, P., 2003. Taking preventive intervention to scale: the Nurse-Family Partnership. Cogn. Behav. Pract. 10, 278–290.

Patterson, G.R., Chamberlain, P., Reid, J.B., 1982. A comparative evaluation of a parent-training program. Behav. Ther. 13, 638-650.

Patterson, G.R., Reid, J.B., Dishion, T.J., 1992. Antisocial Boys. Castalia, Eugene, OR.

Pekarik, E., Prinz, R., Leibert, C., Weintraub, S., Neale, J., 1976. The pupil evaluation inventory: a socio-metric technique for assessing children's social behavior. J. Abnorm. Child Psychol. 4, 483–491.

PL 107-110. No Child Left Behind Act of 2001. 20 USC 6301.

Reid, J.B., 1993. Prevention of conduct disorder before and after school entry: relating interventions to development findings. Dev. Psychopathol. 5, 243-262.

Reid, J.B., Eddy, J.M., 1997. The prevention of antisocial behavior: some considerations in the search for effective interventions. In: Stoff, D.M., Breiling, J., Maser, J.D. (Eds.), Handbook of Antisocial Behavior. John Wiley & Sons, New York, pp. 343-356.

Robins, L.N., 1978. Sturdy childhood predictors of adult antisocial behavior: replications from longitudinal studies. Psychol. Med. 8, 611–622.
Romanoski AJ, Folstein MF, Nestadt G, Chahal R, Merchant A, Brown CH, Gruenberg EM, McHugh PR., 1992. The epidemiology of psychiatrist-ascertained depression and DSM-III depressive disorders. Results from the Eastern Baltimore Mental Health Survey Clinical Reappraisal. Psychol. Med. 22, 629-55.
Sameroff, A., 1994. Developmental systems and family functioning. In: Parke, R.D., Kellam, S.G. (Eds.), Exploring Family Relationships with Other Social Contexts. Vol. 8. Lawrence J. Erlbaum Associates, Hillsdale, NJ, pp. 199–214.
Schafer, J.L., 1997. Analysis of Incomplete Multivariate Data. Chapman & Hall, London.
Schwartzman, A.E., Ledingham, J.E., Serbin, L.A., 1985. Identification of children at-risk for adult schizophrenia: a longitudinal study. Int. Rev. Appl. Psychol. 34, 363–380.
Shedler, J., Block, J., 1991. Adolescent drug use and psychological health: a longitudinal study. Annu. Progr. Child Psychiatr. Child Dev. 24, 545–584.
Spitzer, R.L., Williams, J.B., Gibbon, M., First, M.B., 1992. The Structured Clinical Interview for DSM-III-R (SCID). I: History, rationale, and description. Arch. Gen. Psychiatry 49, 624–629.
Stram, D.A., Lee, H.W., 1994. Variance components testing in the longitudinal mixed effects model. Biometrics 50, 1171–1177.
Turner, R.J., Gil, A.G., 2002. Psychiatric and substance use disorders in south Florida. Arch. Gen. Psychiatry 59, 43-50.
Tingstrom, D.H., Sterling-Turner, H.E., Wilczynski, S.M., 2006. The good behavior game: 1969-2002. Behav. Modif. 30, 225-253.
Vlahov, D., Anthony, J.C., Munoz, A., Margolick, I., Nelson, K.E., Celentano, D., Solomon, L., Polk, B.F., 1991. The ALIVE Study: a longitudinal study of HIV-1 infection in intravenous drug users: description of methods. J. Drug Issues 21, 759–776.
Wade, T.J., Cairney, J., Pevalin, D.J., 2002. Emergence of gender differences in depression during adolescence: national panel results from three countries. J. Am. Acad. Child Adolesc. Psychiatry 41, 190-198.
Wang, Y., 1998. Mixed effects smoothing spline analysis of variance. J. R. Stat. Soc. Ser. B Stat. Methodol. 60, 159–174.
Wang C.P., Brown, C.H., Bandeen-Roche, K., 2005. Residual diagnostics for growth mixture models: examining the impact of a preventive intervention on multiple trajectories of aggressive behavior. J. Am. Stat. Assoc. 100, 1054–1076.
Webster-Stratton, C., 1989. Systematic comparison of consumer satisfaction of three cost-effective parent training programs for conduct-problem children. Behav. Ther. 20, 103–115.
Webster-Stratton, C., 1998. Parent training with low-income clients: promoting parental engagement through a collaborative approach. In: Lutzker, J.R.

(Ed.), Child Abuse: A Handbook of Theory, Research and Treatment. Plenum Press, New York, pp. 183–210.
Weller, E.B, Kloos, A., Kang, J., Weller, R.A., 2006. Depression in children and adolescents; does gender make a difference? Curr. Psychiatry Rep. 8, 108-114.
Werthamer-Larsson, L., Kellam, S.G., Wheeler, L., 1991. Effect of first grade classroom environment on child shy behavior, aggressive behavior, and concentration problems. Am. J. Community Psychol. 19, 585–602.
Wittchen, H.U., 1994. Reliability and validity studies of the WHO Composite International Diagnostic Interview (CIDI): a critical review. J. Psychiatr. Res. 28, 57–84.
Wolfinger, R.D., O'Connell, M., 1993. Generalized linear mixed models: a pseudo-likelihood approach. J. Statist. Comput. Simul. 48, 233–243.
Wood, S.N., 2004. Low Rank Scale Invariant Tensor Product Smooths for Generalized Additive Mixed Models. Technical Report 04-12. Department of Statistics, University of Glasgow, Glasgow, UK.

TABLES

Table 1
GBG and Internal GBG Control comparability on baseline variables (N=407)

Baseline Variable	GBG Means (SD)	Internal GBG Control Means (SD)	*p*-value
Aggressive, disruptive behavior*	2.06 (1.18)	2.01 (0.92)	0.30
Shy behavior*	2.47 (1.03)	2.48 (0.75)	0.43
Concentration problems*	2.89 (1.5)	3.13 (1.14)	0.73
Anxiety**	0.47 (0.2)	0.49 (0.22)	0.17
Depressive symptoms**	0.47 (0.32)	0.60 (0.35)	0.01
Reading achievement†	263.66 (35.88)	259.54 (28.19)	0.84
Mathematics achievement†	302.65 (67.83)	301.37 (60.6)	0.94
Classroom size††	29.75 (5.99)	28.17 (6.68)	0.66
Free or reduced-price lunch status	51%	53%	0.30

*rated by teacher
**self-rated
†standardized achievement rating
†† within schools and across design conditions

Table 2
Presence of Baseline Teacher Ratings and Young Adult Follow-up Status for Children Present in Fall of First Grade, 1985 (*N*=922)

Factor (percent of sample)	n	Teachers' Ratings[4] (%)	Interviewed at Young Adult Follow-up (%)
Overall	922	826 (90%)	689 (75%)
Gender		*p*=0.62	*p*<0.001
Male (50%)	462	412 (89%)	307 (66%)
Female (50%)	460	414 (90%)	382 (83%)
Intervention condition		*P*<0.001	*p*=0.81
Internal GBG control (18%)	169	149 (88%)	126 (75%)
Internal ML control (22%)	205	191 (93%)	153 (75%)
External control (34%)	310	255 (82%)	227 (73%)
GBG (26%)	238	231 (97%)	183 (77%)
Urban area		*p*<0.001	*p*<0.001
Greek/Italian, low/middle class (14%)	126	114 (90%)	82 (65%)
African American, very poor (26%)	243	231 (95%)	187 (77%)
African American, low–middle class (21%)	194	170 (88%)	156 (80%)
African American/white, middle class (23%)	211	173 (82%)	166 (79%)
White, middle class (16%)	148	138 (93%)	98 (66%)
School lunch		*p*=0.54	*p*=0.13
Free or reduced price (47%)	426	379 (89%)	329 (77%)
Not free or reduced price (53%)	490	442 (90%)	357 (73%)

	Mean for followed up (SE)	Mean for attritted (SE)	p-value
Youth baseline measures[5]			
Aggressive, disruptive behavior* (76%)	1.88 (0.04)	1.96 (0.08)	0.32
Shy behavior* (76%)	2.61 (0.04)	2.69 (0.07)	0.38

[4]Fall of 1st Grade, 1985

[5]proportion measured that were followed up in young adulthood

Concentration problems* (76%)	2.93 (0.05)	3.01 (0.10)	0.48
Anxiety** (77%)	0.477 (0.008)	0.497 (0.016)	0.23
Depressive symptoms** (77%)	0.540 (0.014)	0.578 (0.025)	0.20
Reading achievement† (75%)	259.0 (1.4)	256.3 (2.2)	0.31
Math achievement† (75%)	302.6 (2.4)	309.4 (4.4)	0.16

*rated by teacher
**self-rated
†standardized achievement rating

Table 3
GLMM Model for Lifetime Illicit Drug Abuse/Dependence among Males (N=269 students, 31 classrooms)

	Coefficient (SE)	*df*	*t*-value	*p*-value
Fixed Effects				
Intercept	-1.442 (0.396)	237	-3.642	<0.0001
Baseline	1.200 (0.299)	237	4.019	<0.0001
GBG vs. internal GBG control (Tx1)	-0.999 (0.450)	27	-2.221	0.035
External control vs. internal GBG control (Tx2)	-0.211 (0.403)	27	-0.525	0.604
Internal ML control vs. internal GBG control (Tx3)	-0.130 (0.448)	27	-0.291	0.773
	SD	*p*-value[a]		
Random Effects				
Classroom	0.003	0.488		

[a]for testing zero variance

Table 4
Model for Lifetime Illicit Drug Abuse/Dependence among Females (*N*=345 students, 31 Classrooms)

	Coefficient (SE)	*df*	*t*-value	*p*-value
Fixed Effects				
Intercept	-2.858 (0.588)	313	-4.864	<0.0001
Baseline	0.337 (0.467)	313	0.722	0.471
GBG vs. internal GBG control (Tx1)	0.232 (0.637)	27	0.365	0.718
External control vs. internal GBG control (Tx2)	0.697 (0.610)	27	1.142	0.263
Internal ML control vs. internal GBG control (Tx3)	0.581 (0.653)	27	0.890	0.381
	SD	*p*-value[a]		
Random Effects				
Classroom	0.008	0.477		

[a]for testing zero variance

Table 5
GAMM for Lifetime Alcohol Abuse/ Dependence for Males and Females (*N*=621 students, 31 Classrooms)

Type of Effect		Effect (variable name)	Coefficient (Logit)	SE	*z*-value	*p*-value
		Intercept	0.284	0.423	0.671	0.503
Fixed Main Effect						
1	Gender	Female vs. male (gender)	-1.121	0.236	-4.741	0.000
2	Intervention Main Effect (adjusted)	GBG vs. internal GBG controls (Tx1)	-0.699	0.349	-2.006	0.045
3	Contrasts among Control Groups	External controls vs. internal GBG controls (Tx2)	0.112	0.309	0.361	0.719
4		Internal ML controls vs. internal GBG controls (Tx3)	-0.383	0.356	-1.075	0.283
		Nonlinear Terms (variable name)	Effects	*df*	χ^2	*p*-value
Fixed Nonlinear Effects						
5		Baseline aggressive, disruptive behavior	Total baseline	3	7.034	0.070
			Linear (baseline)	1	2.534	0.111
			Smooth (baseline)	2	4.500	0.105
		Effect Name	SD	*p*-value[a]		
Random Effects						
10	Classroom	Classroom level	0.004	0.500		

[a]for testing zero variance

Table 6
GLMM for Males Smoking ≥10 Cigarettes/Day (*N*=278 students)

	Coefficient (SE)	*df*	*t*-value	*p*-value
Main Effects				
Intercept	-4.435 (1.239)	243	-3.579	0.000
Baseline	3.306 (1.144)	243	2.889	0.004
GBG vs. internal GBG control (Tx1)	2.977 (1.441)	27	2.066	0.049
External control vs. internal GBG control (Tx2)	3.234 (1.345)	27	2.404	0.023
Internal ML control vs. internal GBG control (Tx3)	2.713 (1.376)	27	1.972	0.059
Interaction Effects				
Tx1 by baseline	-6.919 (2.094)	243	-3.304	0.001
Tx2 by baseline	-3.848 (1.273)	243	-3.022	0.003
Tx3 by baseline	-2.869 (1.313)	243	-2.184	0.030
	SD	*p*-value[a]		
Random Effects				
Classroom	1.048	0.196		

[a]for testing zero variance

Table 7
GLMM for Females Smoking ≥10 Cigarettes/Day (*N*=348 students)

	Coefficient (SE)	*df*	*t*-value	*p*-value
Main Effects				
Intercept	-4.435 (0.955)	316	-4.646	0.000
Baseline	1.338 (0.479)	316	2.790	0.006
GBG vs. internal GBG control (Tx1)	0.680 (1.133)	27	0.601	0.553
External control vs. internal GBG control (Tx2)	1.465 (1.073)	27	1.365	0.183
Internal ML control vs. internal GBG control (Tx3)	0.705 (1.196)	27	0.589	0.561
	SD	*p*-value[a]		
Random effects				
Classroom	1.562	0.248		

[a]for testing zero variance

Table 8
GAMM for ASPD for Males and Females (N = 629 Students, 31 Classrooms)

	Type of Effect	Effect (Variable Name)	Coefficient (Logit)	SE	*z*-value	*p*-value
		Intercept	0.320	0.894	0.359	0.720
Fixed Main Effects						
1	Gender	Female vs. male (gender)	-0.578	0.435	-1.328	0.185
2	Intervention main effect (adjusted)	GBG vs. internal GBG controls (Tx1)	-2.946	1.026	-2.872	0.004
3	Contrasts among control	External controls vs. internal GBG controls (Tx2)	0.103	0.707	0.146	0.884
4	groups	Internal ML controls vs. internal GBG controls (Tx3)	-0.807	0.729	-1.106	0.269

		Effects	*df*	χ^2	*p*-value
Fixed Nonlinear Effects					
Nonlinear terms (variable name)					
5	Baseline aggressive, disruptive behavior (baseline)	Total baseline	3	7.978	0.046
		Linear (baseline)	1	7.668	0.006
		Smooth (baseline)	2	0.310	0.856
6	Gender * baseline aggressive, disruptive behavior	Total gender * baseline	3	3.222	0.359
		Linear (gender * baseline)	1	1.202	0.272
		Smooth (gender * baseline)	2	2.020	0.364
7	Baseline * GBG vs. internal control (Tx1)	Total Baseline * Tx1	3	9.106	0.028
		Linear(Baseline * Tx1)	1	1.184	0.277
		Smooth (Baseline * Tx1)	2	7.922	0.019
8	Baseline * External vs. internal GBG control (Tx2)	Total baseline * Tx2	3	1.582	0.663
		Linear (baseline * Tx2)	1	0.552	0.458
		Smooth (baseline*Tx2)	2	1.030	0.598
9	Baseline * Internal MLcontrol vs. internal GBG control (Tx3)	Total baseline * Tx3	3	3.292	0.348
		Linear (baseline * Tx3)	1	3.276	0.070
		Smooth (baseline * Tx3)	2	0.016	0.992

		Effect Name	SD	*p*-value[a]
Random Effects				
10	Classroom	Classroom level	0.005	0.500

[a]for testing zero variance

Table 9

GAMM for High School Graduation for Males (N=278 students, 31 Classrooms)

Type of Effect		Effect (Variable Name)	Coefficient (Logit)	SE	*z*-value	*p*-value
		Intercept	0.446	1.090	0.409	0.683
Fixed Main Effects						
Row 1	Poverty	Free lunch indicator	-0.837	0.400	-2.092	0.037
2	Urban region 2	Urban region 2 vs. urban region 1	0.411	0.568	0.724	0.470
3	Urban region 3	Urban region 3 vs. urban region 1	0.674	0.546	1.235	0.218
4	Urban region 4	Urban region 4 vs. urban region 1	1.575	0.532	2.958	0.003
5	Urban region 5	Urban region 5 vs. urban region 1	2.653	0.666	3.984	0.000
2	Intervention main effect (adjusted)	GBG vs. internal GBG controls (Tx1)	0.402	1.332	0.302	0.763
3	Contrasts among control groups	External controls vs. internal GBG controls (Tx2)	-1.593	1.244	-1.281	0.201
4		Internal ML controls vs. internal GBG controls (Tx3)	-0.679	1.234	-0.550	0.583

		Effects	df	χ^2	*p*-value
Fixed Nonlinear Effects					
Nonlinear Terms (Variable Name)					
5	Baseline aggressive, disruptive behavior (baseline)	Total baseline	3	2.074	0.557
		Linear (baseline)	1	0.942	0.332
		Smooth (baseline)	2	1.132	0.568
6	Baseline * GBG vs. internal control (Tx1)	Total baseline * Tx1	3	0.228	0.973
		Linear (baseline * Tx1)	1	0.128	0.721
		Smooth (baseline * Tx1)	2	0.100	0.951
7	Baseline * External vs. Internal GBG control (Tx2)	Total baseline * Tx2	3	1.994	0.574
		Linear (baseline * Tx2)	1	0.032	0.858
		Smooth (baseline * Tx2)	2	1.962	0.375
8	Baseline * Internal MLcontrol vs. Internal GBG control (Tx3)	Total baseline * Tx3	3	1.024	0.795
		Linear (Baseline * Tx3)	1	0.136	0.712
		Smooth(Baseline*Tx3)	2	0.888	0.641

		Effect Name	SD	*p*-value[a]
Random Effects				
9	Classroom	Classroom level	0.007	0.500

[a]for testing zero variance

Table 10
Comparison of Rates of Young Adult Outcomes for GBG and Internal Controls

Outcome	GBG, %	Controls, %	Relative Risk for GBG Compared With Controls, %	*p*-value
Lifetime illicit drug use				
Males	19	38	50	0.035
Lifetime alcohol abuse/dependence				
All	13	20	65	0.04
Smoking ≥10 cigarettes/day				
Males	7	17	41	0.05
ASPD				
All	17	25	68	0.07
Highly Aggressive, disruptive	41	86	48	0.02

FIGURE LEGENDS

Fig. 1. Impact of GBG vs. all 3 controls on drug abuse/dependence disorder by baseline aggressive, disruptive behavior for males and females.

Fig. 2. Impact of GBG vs. all 3 control groups combined on drug abuse/dependence disorder by baseline aggressive, disruptive behavior for males.

Fig. 3. Impact of GBG vs. all 3 control groups combined on drug abuse/dependence disorder by baseline aggressive, disruptive behavior for females.

Fig. 4. Impact of GBG vs. all 3 controls on alcohol abuse/dependence disorder by baseline aggressive, disruptive behavior for males and females.

Fig. 5. Impact of GBG vs. all 3 controls on regular smoking by baseline aggressive, disruptive behavior for males.

Fig. 6. Impact of GBG vs. all 3 controls on ASPD by baseline aggressive, disruptive behavior for males and females.

Fig. 7. Impact of GBG vs. all 3 control groups on completed high school by baseline aggressive, disruptive behavior for males.

Fig. 1.

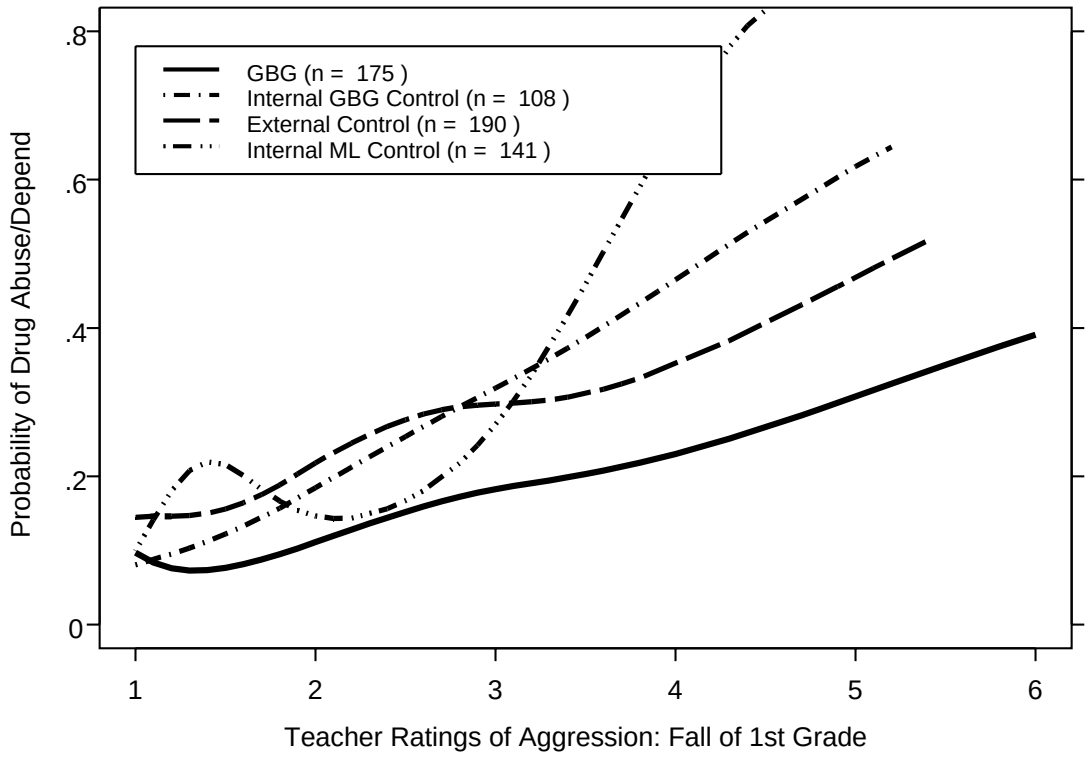

GBG (n = 175)
Internal GBG Control (n = 108)
External Control (n = 190)
Internal ML Control (n = 141)
Probability of Drug Abuse/Depend
0
.2
.4
.6
.8
1
2
3
4
5
6
Teacher Ratings of Aggression: Fall of 1st Grade

Fig. 2.

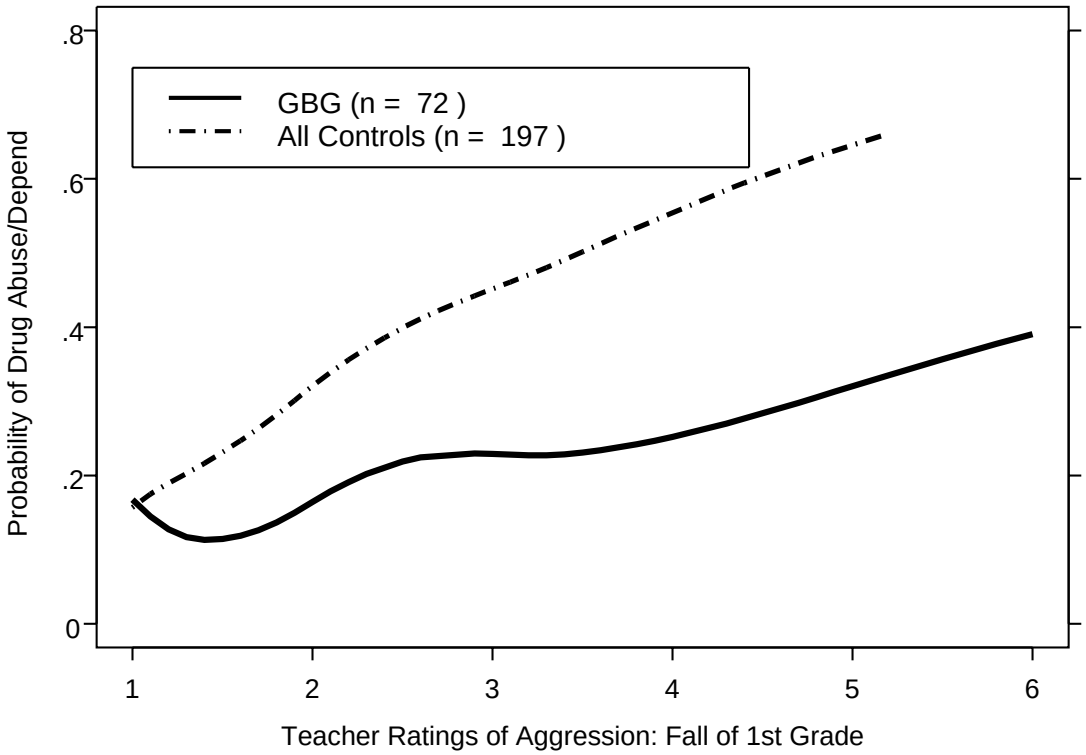

GBG (n = 72)
All Controls (n = 197)
Probability of Drug Abuse/Depend
0
.2
.4
.6
.8
1
2
3
4
5
6
Teacher Ratings of Aggression: Fall of 1st Grade

Fig. 3.

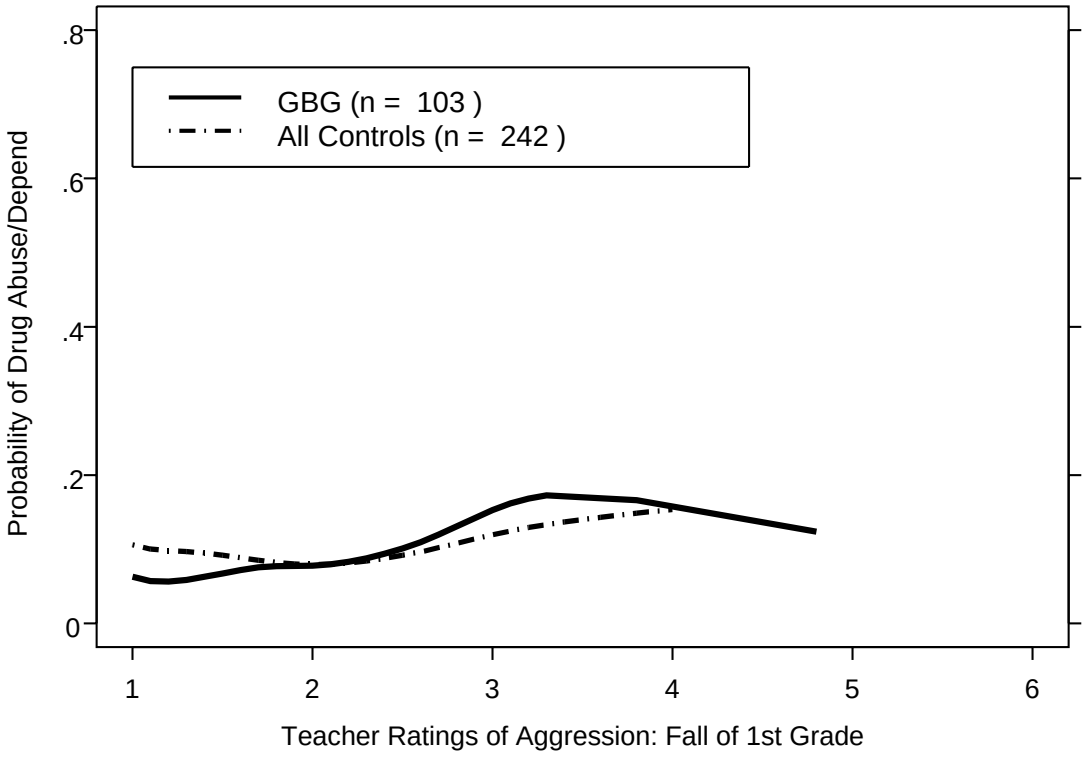

GBG (n = 103)
All Controls (n = 242)
Probability of Drug Abuse/Depend
0
.2
.4
.6
.8
1
2
3
4
5
6
Teacher Ratings of Aggression: Fall of 1st Grade

Fig. 4.

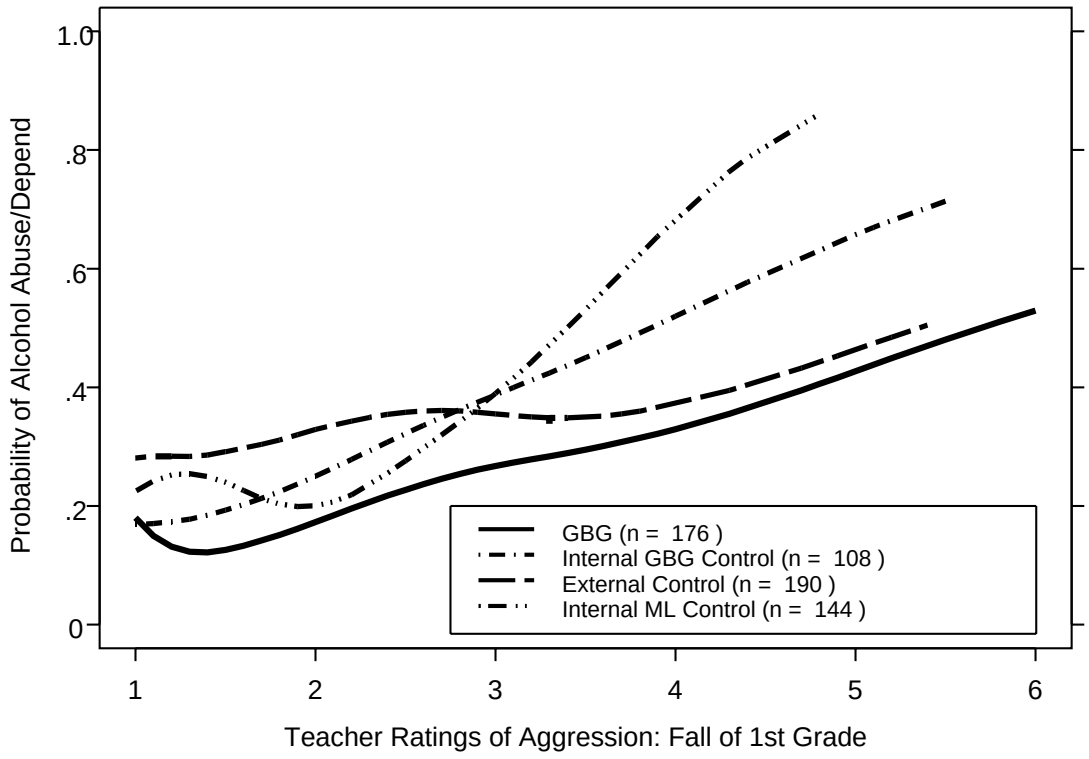

Probability of Alcohol Abuse/Depend
1.0
.8
.6
.4
.2
0
1
2
3
4
5
6
Teacher Ratings of Aggression: Fall of 1st Grade
GBG (n = 176)
Internal GBG Control (n = 108)
External Control (n = 190)
Internal ML Control (n = 144)

Fig. 5.

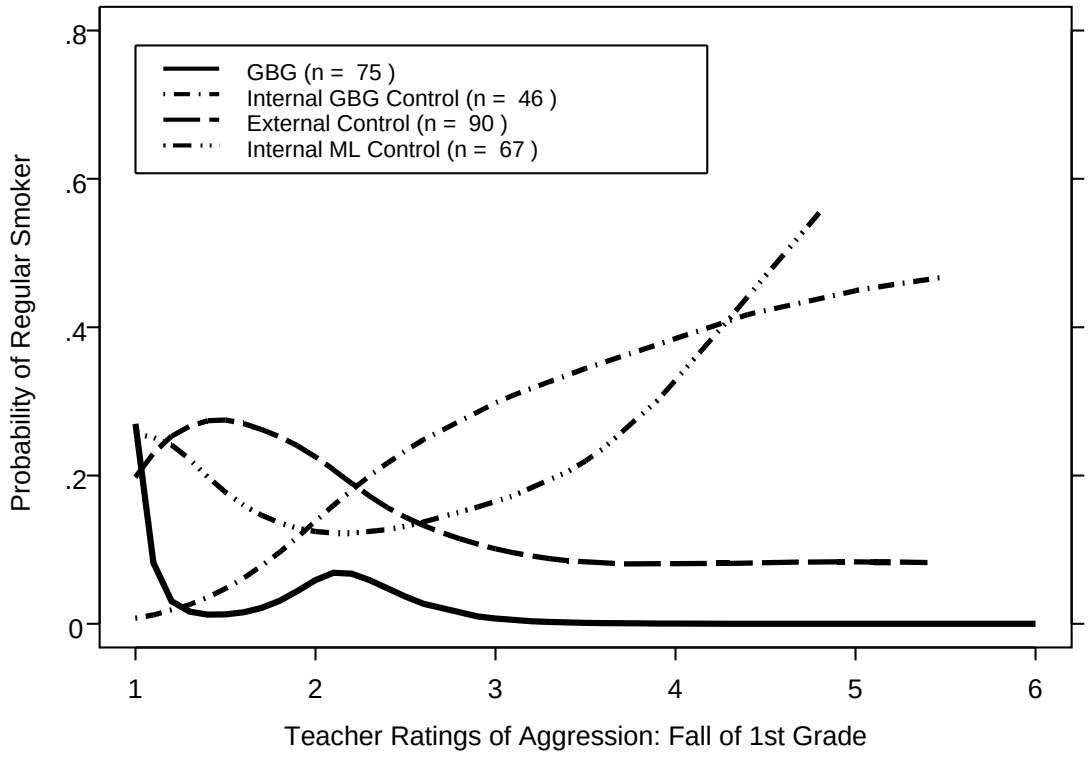

GBG (n = 75)
Internal GBG Control (n = 46)
External Control (n = 90)
Internal ML Control (n = 67)
Probability of Regular Smoker
.8
.6
.4
.2
0
1
2
3
4
5
6
Teacher Ratings of Aggression: Fall of 1st Grade

Fig. 6.

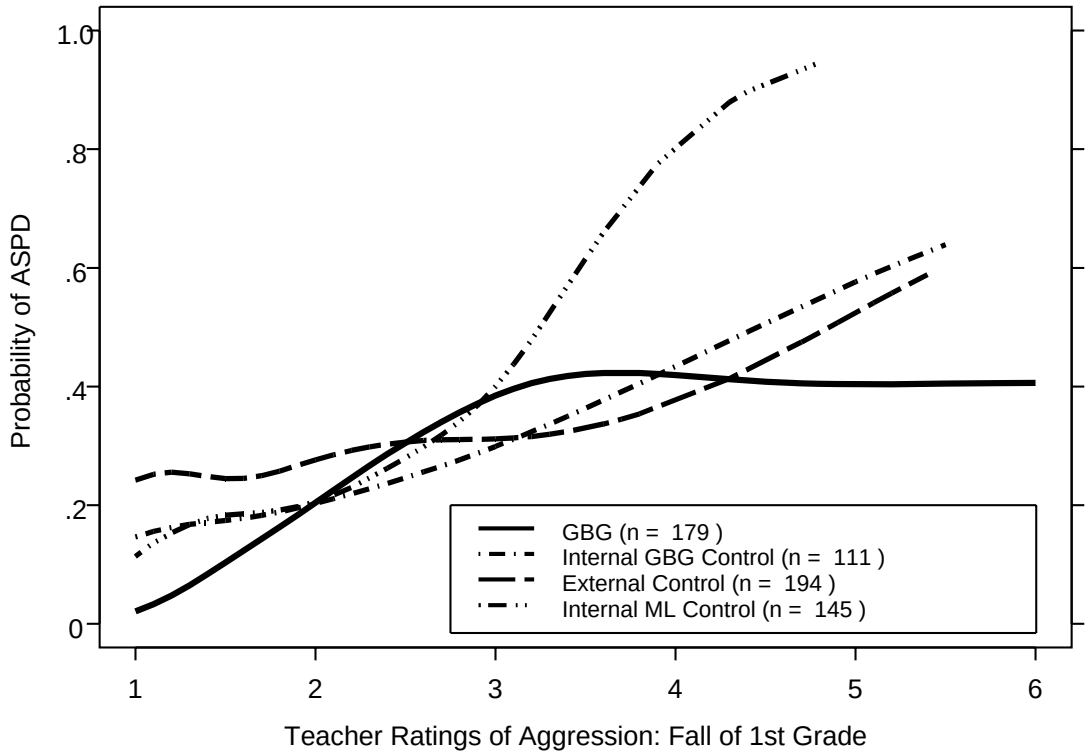
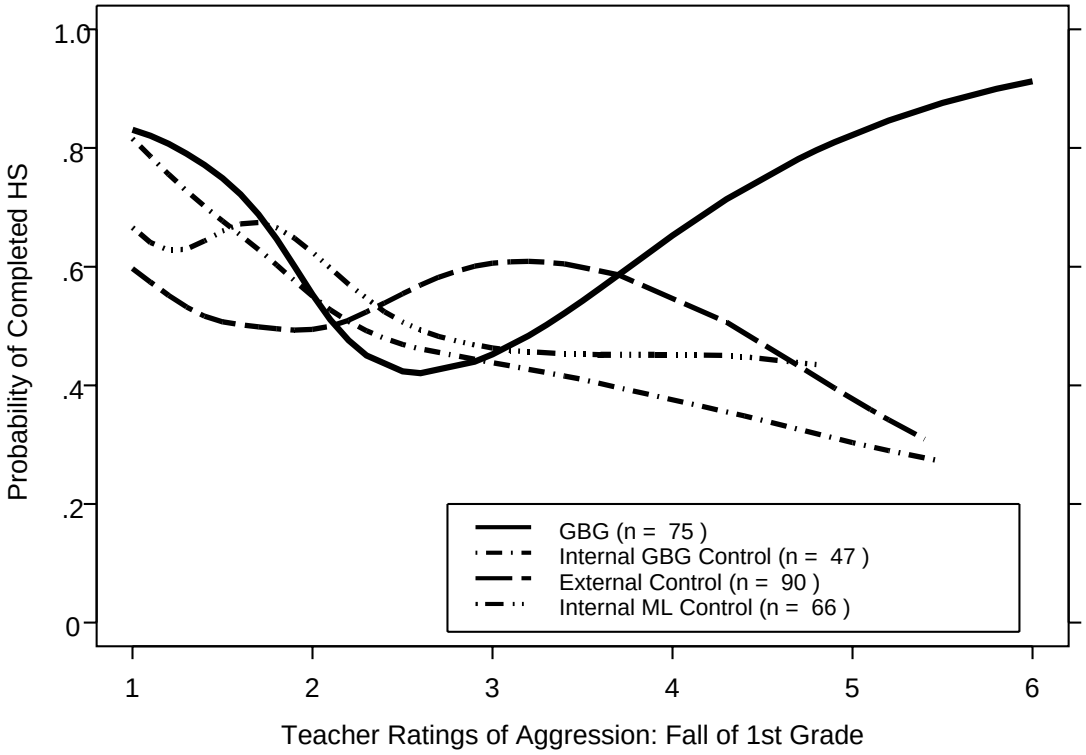

1.0
.8
.6
.4
.2
0
Probability of ASPD
1
2
3
4
5
6
Teacher Ratings of Aggression: Fall of 1st Grade
GBG (n = 179)
Internal GBG Control (n = 111)
External Control (n = 194)
Internal ML Control (n = 145)

Fig. 7


1.0
.8
.6
.4
.2
0
Probability of Completed HS
1
2
3
4
5
6
Teacher Ratings of Aggression: Fall of 1st Grade
GBG (n = 75)
Internal GBG Control (n = 47)
External Control (n = 90)
Internal ML Control (n = 66)