Published in final edited form as:

Soc Serv Rev. 2010 ; 84(4): 615–655.

Dosage Effects on School Readiness: Evidence from a Randomized Classroom-Based Intervention

Fuhua Zhai,

Stony Brook University

C. Cybele Raver,

New York University

Stephanie M. Jones,

Harvard University

Christine P. Li-Grining,

Loyola University Chicago

Emily Pressler, and

Pennsylvania State University

Qin Gao

Fordham University

Abstract

Variations in the dosage of social interventions and the effects of dosage on program outcomes remain understudied. This study examines the dosage effects of the Chicago School Readiness Project, a randomized, multifaceted classroom-based intervention conducted in Head Start settings. Using a principal score matching method to address the issue of selection bias, the study finds that high-dosage levels of teacher training and mental health consultant class visits have larger effects on children's school readiness than the effects estimated through intention-to-treat (ITT) analyses. Low-dosage levels of treatment are found to have effects that are smaller than those estimated in ITT analyses or to have no statistically significant program effects. Moreover, individual mental health consultation services provided to high-risk children are found to have statistically significant effects on their school readiness. The study discusses the implications of these findings for research and policy.

Research suggests that program participation is a critical source of variance in the outcomes associated with early childhood interventions. Exposure to different levels of services provided (hereafter described as dosage) can reveal whether such interventions yield statistically significant benefits (Shonkoff and Phillips 2000; Hill, Brooks-Gunn, and Waldfogel 2003). Investigators favor randomized experiments as the gold standard for evaluation of programs in which treatment and control groups have similar observed and unobserved characteristics (Hill et al. 2003; Peck 2003; Agodini and Dynarski 2004; Smith and Todd 2005; Lochman et al. 2006). Nevertheless, due to participants' noncompliance with randomized assignments, most experimental interventions involving human subjects suffer from complications (Barnard et al. 2003; Peck 2003). Moreover, many social

interventions provide multifaceted program services, and evaluation can be difficult because, in addition to the issue of noncompliance, participants are also exposed to different subcomponents of these interventions, which make it difficult to isolate the respective individual effects of the various subcomponents. The analyses of differentials in participants' dosage levels may have serious implications for policy makers and program administrators. However, the empirical issue of self-selection limits the study of those matters because, after random assignment, individual members of the treatment group can make choices about whether to receive, and the extent to which they receive, the intervention services.

To address the issues of noncompliance and self-selection existing in social experiments, the current study investigates the dosage effects of an early childhood intervention program. Specifically, it employs data from the Chicago School Readiness Project (CSRP), a randomized, multifaceted classroom-based intervention conducted in Head Start settings. These data are analyzed to examine the effects of differential dosage levels of the CSRP intervention on children's school readiness. The study also considers whether those dosage effects vary by subcomponents of the CSRP intervention. The study employs innovative analytic methods to address the issue of selection bias.

Background and Prior Research

Preschool is an important setting for early childhood education outside the home. Approximately 60 percent of 4-year-olds in the United States enroll in some form of preschool prior to kindergarten (Webster-Stratton and Taylor 2001; Waldfogel 2006). Substantial research suggests that preschool is an effective setting for preventive interventions designed to reduce children's behavioral problems and to promote their cognitive development. Findings also suggest that such interventions are particularly effective for economically disadvantaged children who face high risk of developing social-emotional difficulties and have limited access to mental health services (Raver 2002; Lochman and Wells 2003; Webster-Stratton, Reid, and Hammond 2004; Smolkowski et al. 2005; Wong et al. 2008).

Some studies identify benefits of classroom-based interventions that focus on providing preschool teachers with training and consultations with mental health professionals. These interventions are found to improve classroom management strategies, overall child-care quality, the quality of teacher-child interactions, teachers' self-efficacy, and their competence in dealing with difficult children (Alkon, Ramler, and Mac-Lennan 2003; Webster-Stratton et al. 2004). Such improvements in teachers' skills are found, in turn, to benefit the development of low-income children.

For instance, recent randomized experiments and quasi-experiments adopt the Incredible Years Training Series, a classroom-based multifaceted intervention designed to prevent and treat behavior problems in young children (ages 3–8). These studies find that training preschool teachers in classroom management strategies is associated with reductions in children's disruptive behaviors. Classroom-based mental health consultation, designed to facilitate the implementation of these strategies, is found to be positively related to

children's improvement in social competence and adaptation (Webster-Stratton, Reid, and Hammond 2001; Reid, Webster-Stratton, and Hammond 2003; Shernoff and Kratochwill 2007; Williford and Shelton 2008). Moreover, research also suggests that individualized, child-focused mental health consultation statistically significantly increases social skills and reduces behavior problems among preschool children with social-emotional difficulties. Such consultation also reduces the probability that a child will be expelled from a child-care center (Perry et al. 2008).

Findings from such randomized controlled trials represent important steps forward for the fields of policy analysis and prevention science, but most of these recent studies only provide intention-to-treat (ITT) estimates for the average treatment effects of classroom-based interventions. As a conventional and rigorous test of program effects, ITT analysis compares the average outcomes of the treatment group with those of the control group. However, the comparison does not consider whether or not participants actually comply with assigned intervention conditions or to what extent they comply (Gibson 2003; Peck 2003; Angrist 2006; Lochman et al. 2006). Thus, ITT analysis has important policy relevance, as it provides estimates of program effects on the population for which the intervention is intended (Webster-Stratton et al. 2001; Peck 2003).

Nevertheless, another critical policy issue to consider is that real-world conditions of program implementation often differ from the aspirations of researchers and policy makers. Many participants do not comply with their assigned treatment conditions and do not take up or fully receive the intervention services (see review by Peck 2003). For example, studies on parenting training programs in preschool settings, including those on programs adopting the Incredible Years Training Series, report that about one-third of parents assigned to the treatment group never attend a session; from 12 percent to more than 50 percent of parents attend less than half of the sessions (Barkley et al. 2000; Webster-Stratton et al. 2001). As a result, ITT estimates may be substantially lower than the actual effects of interventions if participants were to fully comply with the treatment assignments, and the ITT estimates may not even be statistically significant in some circumstances in which the intervention potentially is effective (Barkley et al. 2000; Gibson 2003; Hill et al. 2003; Angrist 2006; Lochman et al. 2006). In addition, many social interventions, such as the Incredible Years Training Series, offer multifaceted services. However, few studies estimate the program effects of these individual components, and there is little guidance on best approaches that might be taken to empirically model the benefits of subcomponent services when they are bundled together in a single package of treatment (Gibson 2003; Peck 2003).

Answers to these empirical questions on program dosage effects have pressing policy relevance. Budget constraints often force resource allocation decisions that attempt to balance program intensity or duration with the number of participants served. If a higher dosage level is found to be associated with better program outcomes, policy makers may be motivated to increase program intensity or duration rather than the number of families or individuals served (Lu et al. 2001). For example, some descriptive evidence suggests that the duration of preschool teachers' training and consultations with mental health professionals predicts both teachers' self-efficacy and child outcomes (Alkon et al. 2003; Green et al. 2006; Brennan et al. 2008). Conversely, if intervention program outcomes resulting from

high-dosage levels are similar to those from low-dosage levels, the marginal returns from high dosage may be relatively low or negligible. As a result, policy makers may consider allocating resources to focus on the number of individuals served rather than on the intensity or duration of existing interventions.

In addition, the issue of self-selection is a major impediment to the investigation of program dosage effects. Because participants can choose whether to take part in treatment (or its subcomponents) after random assignment, and the extent to which they participate, the dosage level of program participation is often associated with the pretreatment characteristics of individual members of the treatment group (Hill, Waldfogel, and Brooks-Gunn 2002; Gibson 2003; Hill et al. 2003; Peck 2003; Lochman et al. 2006). For example, studies on parenting training programs in preschool settings report that treatment group parents who do not attend or drop out of the programs tend to be younger, at lower family risk, and less educated than those who attend or continue in the interventions; their children also tend to have fewer behavior problems at baseline (Barkley et al. 2000; Webster-Stratton et al. 2001). These pretreatment covariates are also associated with children's developmental outcomes (Brooks-Gunn, Duncan, and Aber 1997; Li-Grining et al. 2006; Zhai, Brooks-Gunn, and Waldfogel, forthcoming). As a result, simply comparing the outcomes of individuals who receive different dosage levels of interventions may lead to biased estimates of dosage effects.

Modeling the role of dosage is additionally complicated in the context of randomized trials, because individuals in the control group are generally not given the option to enroll in the treatment, and their level of participation (or their dosage levels) therefore cannot be observed directly. Nevertheless, evidence from the treatment group in a randomized experiment, as the discussion above mentions, suggests that only a subgroup of the control-group-assigned participants would actually take up or fully complete the program if it were offered to them (Frangakis and Rubin 2002; Gibson 2003; Peck 2003). As a consequence, biased estimates of dosage effects may result from simple comparisons of the outcomes of individuals with different dosage levels, including those in the entire control group (Bickman, Andrade, and Lambert 2002; Hill et al. 2003; Peck 2003; Lochman et al. 2006).

Recently, several methods have been designed to estimate program dosage effects and to address the issue of self-selection. These methods use either matching or weighting strategies. Two of them directly model the dosage of treatment. One such method uses a single scalar balancing score estimated by an ordered logistic regression for multiple dosage levels. It then conducts nonbipartite pair matching (i.e., in contrast to bipartite matching between treatment and control groups) based on the estimated distance scores (Joffe and Rosenbaum 1999; Lu et al. 2001; Guo and Fraser 2009). The goal of this method is to identify pairs of cases that are similar in terms of observed covariates but very different in terms of dosage levels. Another approach uses a multinomial logit model to estimate multiple balancing scores. The inverse of a specific, generalized propensity score is then used as sampling weight to conduct an outcome analysis (i.e., analysis with propensity score weighting; Imbens 2000; Guo and Fraser 2009).

A third method, matching estimators, is widely used in evaluation research. This approach does not model treatment dosage directly and does not use logistic regressions to predict propensity scores. Instead, it uses a vector norm (calculated by either the inverse of the sample variance matrix or the inverse of the sample variance-covariance matrix) to estimate the distances on observed covariates between treated and control cases. It then estimates the differences in outcomes between the treated and control cases that have the shortest distances (Abadie and Imbens 2002, 2006; Guo and Fraser 2009). Since the approach of matching estimators imputes a potential outcome for each study unit, it is possible to estimate the average treatment effects for user-defined subsets of units. For example, Shenyang Guo and Mark Fraser (2009) employ the matching estimators method to evaluate the treatment effects of program dosage in a randomized intervention that involved a skills-training curriculum for elementary school students in North Carolina.

These evaluation methods offer a number of methodological advantages. Research suggests that they are effective and robust tools for the evaluation of the treatment effects of program dosage (Imbens 2000; Lu et al. 2001; Guo and Fraser 2009). However, these methods may not be directly applicable to the analysis of data from relatively small randomized trials that involve multilevel treatment units, as is the case in the current study.

In particular, these methods that directly model dosage of treatment use either a single scalar balancing score estimated by an ordered logistic regression or multiple balancing scores estimated by a multinomial logit model. Results from observational studies suggest that, because all individuals in both treatment and control groups are exposed to treatment, their dosage levels of participation may vary from no dosage to high levels; they also may reflect other defined levels (e.g., Imbens 2000; Lu et al. 2001; Guo and Fraser 2009). In contrast, control group participants in randomized interventions are not exposed to treatment. However, one should not assume that the potential dosage levels would be zero for all control group members. The dosage levels in the control group would also vary in a similar way as those in the treatment group if control group participants had been randomly assigned to the treatment condition. Therefore, it would be inappropriate to directly conduct dosage analyses using ordered or multinomial logistic models on the full sample (i.e., including both treatment and control groups) of a randomized intervention.

In addition, the current study's small sample limits the ways in which dosage levels can be defined for participants in the treatment group. This study therefore relies on a dichotomized measure (i.e., low or high dosage), rather than on use of a scale with multiple dosage levels, to empirically characterize participation. This limitation also prevents this study from using ordered or multinomial logistic models to conduct dosage analyses with only the treated group. Matching estimators is a promising method for use in dosage analyses because it is intuitively appealing and easy to implement (Abadie and Imbens 2002, 2006; Guo and Fraser 2009). However, this study must also consider the issue of clustering, because the matching estimators method does not correct for inefficiency induced by clustering (Guo and Fraser 2009). The current version of the matching estimators approach is not suitable for this study's dosage analyses, because the study relies on data from a clustered randomized trial with multilevel treatment units.

Principal stratification and subgroup analysis provide an alternative means to address the issue of selection bias in studying the dosage effects of multifaceted social interventions (Frangakis and Rubin 2002; Hill et al. 2002, 2003; Barnard et al. 2003; Gibson 2003; Peck 2003). A few studies examine the dosage effects of interventions on young children's developmental outcomes by employing such analytic approaches as principal score matching, propensity score matching, complier average causal effect models, and instrumental variables estimation. Most of these studies find that dosage levels are positively associated with intervention effects. For example, some analyses find no ITT effects but statistically significant treatment effects among minimal-level compliers (Lochman et al. 2006). In addition, some estimates suggest that the relation between dosage levels and intervention effects is larger and longer lasting among high-level compliers than among their minimally compliant counterparts (Angold et al. 2000; Hill et al. 2003).

The Present Study

This study adopts a principal score matching method to address the issue of selection bias in program dosage analyses. It investigates the dosage effects of the CSRP intervention, examining three key indicators of children's school readiness: teacher-reported behavior problems, observer-reported emotional and behavioral self-regulation skills, and observer-rated cognitive development. In addition, this study examines whether the dosage effects vary across school readiness measures and the individual components of the CSRP intervention.

Previous research finds that the CSRP has statistically significant ITT effects on children's school readiness outcomes. These effects include reductions of children's behavior problems as well as improvements in social-emotional skills and cognitive development (Raver et al. 2009; Raver et al., forthcoming). These findings and results from evaluations of other interventions that target young children at risk lead the authors to identify four hypotheses. First, this study posits that the estimated treatment effects for children who received high doses of the CSRP intervention will be larger than corresponding ITT estimates. Second, it posits that the estimated effects for children with low doses of treatment will be smaller than ITT estimates or that the estimates will be statistically nonsignificant. Third, the dosage effects are posited to differ across measures of school readiness. Fourth, this study hypothesizes that dosage effects will vary across individual components of the CSRP intervention.

Method

CSRP Participants and Intervention

The CSRP focuses on teachers of ethnic minority preschoolers, on enhancing the quality of emotional support provided by these teachers, and on helping them to build effective classroom management strategies. These efforts are intended to support children's development of self-regulation, reduce their risk of behavioral difficulty, and increase their opportunities for learning (Raver et al. 2009). The CSRP randomly assigns two cohorts of Head Start program children and teachers into a multifaceted classroom-based intervention. These two cohorts come from seven of the most economically disadvantaged neighborhoods

in Chicago. Cohort 1 participated in the intervention from fall 2004 to spring 2005. Cohort 2 participated from fall 2005 to spring 2006. Data for each cohort were collected at two points: September (pretreatment) and May (posttreatment) of the Head Start year.

The CSRP adopts a clustered, randomized controlled trial (RCT) design and a pairwise matching procedure (Bloom 2005) to recruit participants and assign them into treatment and control groups. In particular, the study first chooses 18 Head Start sites that have two or more classrooms and provide full-day services in the seven economically disadvantaged Chicago neighborhoods. The 18 sites are then matched into nine pairs based on a range of site-level demographic characteristics that each site collects and reports annually to the federal government. One site in each matched pair is then randomly assigned to the treatment group, and the other to the control group. Each site initially included two randomly selected classrooms. After randomization, one classroom left the study due to Head Start funding cuts. In total, CSRP participants include 602 children and 90 teachers in 35 classrooms at 18 Head Start sites.

The data collected in September of the Head Start year using questionnaires for parents and teachers show that, on average, participating children in the CSRP are 4 years old, and about half are boys. Approximately 66 percent of participating children are non-Hispanic black, 26 percent are Hispanic, and 8 percent are members of other racial or ethnic groups. On average, sampled teachers reported that they were 40 years old at the cohort's baseline interview (September of the Head Start year), and almost all teachers (97 percent) are female. About 70 percent of teachers identify themselves as non-Hispanic black, 20 percent identify as Hispanic, and 10 percent identify themselves as non-Hispanic white.

The CSRP intervention in the treatment group includes three service components. The first component is a 30-hour teacher training that focuses on behavior management strategies. These strategies are adapted from the Incredible Years teacher training module (Webster-Stratton et al. 2004). All treatment-assigned teachers, including head teachers and assistant teachers, were invited to participate in the five 6-hour training sessions, which were held on Saturdays from September to March during the Head Start year. The participation was voluntary. The incentives for teachers to participate in the training sessions included a compensative payment at a rate of \$15 per hour, catered lunches, and on-site child care.

The second component of the CSRP intervention is the placement of mental health consultants (MHCs) into treatment group classrooms. These clinically trained MHCs held master's degrees in social work and had experience working with young children from low-income families. Attending classes one morning per week, these MHCs coach teachers in implementing behavior management strategies and support them in the use of specific techniques to promote children's positive emotion and behavioral development. They also provide ongoing stress management consultation to help teachers deal with work-related stress. The stress reduction strategies that the MHCs provide are tailored to teachers' individual needs.

The third component of the CSRP intervention involves individual mental health consultation services for a small number of children (three to four children per class) in the

treatment group. Children who receive these services have high emotional and behavioral problems, as identified by the MHCs based on their clinical judgment, consultation with teachers, and review of teacher-reported measures of children's behavioral problems in September of the Head Start year. These children receive individualized, direct intervention services from the MHCs, including individual and group therapies. The services were provided from March to May in the Head Start year.

To ensure that the child-staff ratio is similar across treatment and control classrooms, the CSRP provided an aide to teachers in the control group. These teachers' aides only provided staffing support during everyday classroom activities and were present in the control group classrooms for the same amount of time per week as the MHCs were in the treatment group classrooms.

Measures

Outcome variables—Children's school readiness is the primary outcome measured in this study. It is measured by behavioral, social-emotional, and cognitive scales completed by teachers and independent observers. As previous research suggests, teachers' reports concerning child outcomes may be biased by knowledge of children's treatment status (Barkley et al. 2000). The study therefore adopts scales completed by both teachers and observers. Observers are graduate students and full-time research staff who have at least a bachelor's degree. They are not informed of the treatment status of the subjects to whom they are assigned. Observers conduct one-on-one assessment of children in their schools.

The measures of school readiness include teacher-reported behavior problems, observer-reported social-emotional skills, and observer-rated cognitive development. These scales provide standardized measures of young children's school readiness. They are used extensively in large-scale policy evaluations as well as in smaller efficacy trials of educational and clinical interventions (e.g., Hill et al. 2002, 2003; Love et al. 2002; Yeung, Linver, and Brooks-Gunn 2002; Puma et al. 2005; Spencer et al. 2005; Hooper and Bell 2006; Markowitz et al. 2006; Cathers-Schiffman and Thompson 2007).

Children's behavior problems are measured by the Behavior Problem Index (BPI). The BPI captures teachers' responses to questions adapted from a 28-item rating scale originally designed for parents to report the types of child behavior problems (Zill 1990). The study uses the scales for internalizing and externalizing problems. Following recommendations from the National Longitudinal Survey of Youth (Zill 1990), the authors sum items to form two domains: internalizing ($\alpha = .80$) and externalizing ($\alpha = .92$).

Children's social-emotional skills are reported by observers using the Preschool Self Regulation Assessment-Assessor Report (Smith-Donald et al. 2007). This measure includes 28 items and is adapted from the Leiter-R social-emotional rating scale (Roid and Miller 1997) as well as from the Disruptive Behavior-Diagnostic Observation Schedule coding system (Wakschlag et al. 2005). The measure provides a global picture of children's emotions, attention, and behavior throughout assessor-child interaction. Data obtained through the social-emotional skills measure are aggregated into two subscales: attention and impulse control ($\alpha = .93$) and positive emotion ($\alpha = .84$). The attention and impulse control

subscale measures whether children pay attention during instructions and demonstrations and whether they think and plan before beginning each task. The positive emotion subscale measures whether children are interactive and show pleasure and confidence during activities.

Children's cognitive development is reported by observers using the third edition of the Peabody Picture Vocabulary Test (PPVT) and a measure of early math skills. A shortened version of the PPVT, using a 24-item scale ($\alpha = .78$), asks children to identify one out of four pictures that correspond to the word or action indicated by the observer (Dunn and Dunn 1997; Zill 2003*b*). The early math skills measure ($\alpha = .82$) consists of 19 items reported by observers that cover basic addition and subtraction (Zill 2003*a*).

Dosage measures—This study relies on three measures of dosage: the hours of teacher training, the hours of MHC class visits, and whether children received individual mental health consultation services. The measures correspond to the three components of the CSRP intervention. Not all teachers in the treatment group completed the 30-hour training sessions. Figure 1 shows the average hours of training received by each of 18 CSRP treatment group classes (i.e., the per-class average reflects hours of training received by all teachers in that class). On average, teachers in each class received 18 hours of training (standard deviation [SD] is 9 hours). As figure 1 shows, all of the teachers in one treatment group class did not attend any training session; teachers in the remaining treatment group classrooms received between 9 and 30 hours of training.

Prior research on the Incredible Years Training Series suggests that attendance of less than half of an intervention is an inadequate dosage level (Webster-Stratton et al. 2001). Figure 1 suggests that, among CSRP teachers who received any training, there are two distinct groups. One group represents teachers in classes for which the hours of received training are concentrated right below the mean. On average, teachers in these classrooms received 9–18 hours of training. Hours received by this group account for 30–60 percent of the total training hours provided by the CSRP. The second group represents classes with teachers who received all or nearly all of the 30 hours of training provided by CSRP. On average, classes in this group received 26–30 hours of training. This accounts for between 87 and 100 percent of the total training hours provided by CSRP. The 8-hour gap (27 percent of the total training) between the ranges of average teacher training hours in these two groups is considerable. The first group is therefore defined as the low-dosage training group. This group represents 28 teachers in 10 classes. The classes serve 175 children. The second is defined as the high-dosage training group. It represents 18 teachers in seven classes; these teachers serve 114 children.

Similarly, the number of hours that MHCs visit classes ranges from 100 to 152 hours per class (the mean is 128 hours; SD is 18 hours), and two distinct groups are discernible: MHC visits of 100–126 hours and visits of 132–152 hours. For this study, the former group is defined as the low-dosage MHC group. It comprises 123 children in eight classrooms. The second group, identified by MHC visits between 132 and 152 hours, is designated the high-dosage MHC group. It comprises 185 children in 10 classrooms. The estimates suggest that the two groups are separated by a 6-hour gap between their ranges of MHC visits.

The third dosage measure is a dichotomous indicator of whether or not a child received any individual mental health consultation services provided by the MHCs. Previous research suggests that the number of hours of mental health consultation per child in early childhood settings is not statistically significantly associated with reductions in children's behavior problems or with improvement in social competence (Green et al. 2006). Only a small number of children ($n = 137$) in the CSRP treatment group received any individual mental health consultation services. The study therefore treats exposure to individual mental health consultation services as an extra dose of the CSRP intervention, in addition to the services of teacher training and MHC class visits, for this group of children who have high emotional and behavioral problems.

The three dosage measures are relatively independent of each other and have very weak correlations. For example, the correlation coefficient (r) between high-dosage teacher training and high-dosage MHC class visits is estimated to be 0.07. The coefficient for the correlation between high-dosage teacher training and receipt of individual mental health consultation services is 0.01. The coefficient for the correlation between high-dosage MHC class visits and individual mental health consultation services is 0.09. Therefore, these dosage measures can be investigated separately to estimate their respective effects on children's outcomes under the assumptions that participants also received the services of other components of the CSRP intervention and that those services were distributed equally or randomly. It should be noted that the findings on the dosage effects of individual components should be interpreted in the context of the CSRP's multifaceted services. Because all children received the treatment of teacher training and MHC class visits and some children also had individual mental health consultation services, the dosage effects of individual components reported in this study might be larger than those of individual components if the components were provided separately.

Baseline covariates—The covariates in this study include the characteristics of children, teachers, classrooms, and sites at the time of the baseline interview (i.e., September of the Head Start year). Results from many preventive interventions targeting low-income children suggest the importance of disaggregating the potential confounders of child and family demographic characteristics (see, e.g., Aber, Brown, and Jones 2003; Tolan, Gorman-Smith, and Henry 2004; Schaeffer et al. 2006). In this study, child-level covariates include child gender, race or ethnicity, poverty-related family risks, whether the child is from a single-parent family, whether his or her parents speak Spanish during the CSRP data collection, and children's pretreatment scores in September for corresponding outcome variables in May. Child race or ethnicity is coded as whether the child is non-Hispanic black, since the majority of children in the CSRP are reported to be either non-Hispanic black (66 percent) or Hispanic (26 percent). The measure of poverty-related family risks is a sum of three risk factors: whether the mother holds less than a high school diploma, whether family income-to-needs ratio is less than half the federal poverty threshold, and whether the mother works 10 or fewer hours per week. Previous analyses with large, nationally representative data sets suggest that these covariates represent the most reduced and informative set of indicators for families' exposure to deep poverty (Raver, Garner, and Smith-Donald 2007). This study

employs a multiple imputation method to address the issue of missing data on child-level covariates. The procedure is detailed in the appendix.

In addition, recent research suggests that the ways in which interventions work may differ with institutional resources and teacher motivation (Gottfredson, Jones, and Gore 2002). Thus, the current analyses include a set of teacher- and class-level covariates that are intended to function as proxies for teaching as well as for classroom quality and environment. Specifically, teachers' personal stressors are indexed by six self-reported risks. These risks include whether they have less than an associate's degree, 3 or fewer years of preschool teaching experience, and depressive symptoms, as well as whether they are single, the primary income earner in their household, or live with four or more minors. Data on teachers' personal stressors are collected through a questionnaire derived from the Cornell Early Social Development Study (Raver 2003). In addition, four items examine teachers' work-related stressors. The assessed stressors include feelings of high job demand, lack of confidence in managing classrooms, low job control, and low job resources. Three of the stressors (i.e., job demand, job control, and job resources) were based on the Child Care Worker Job Stress Inventory (CCW-JSI; Curbow et al. 2000). The CCW-JSI is designed to capture child-care providers' self-reported assessments of job stress. Another work-related stressor (i.e., lack of confidence) is adopted to measure teachers' beliefs regarding the causes of children's behavior as well as their confidence in handling misbehavior (Scott-Little and Holloway 1992; Hammarberg and Hagekull 2002). Classroom quality is assessed using the Classroom Assessment Scoring System (CLASS; La Paro, Pianta, and Stuhlman 2004) and the revised edition of the Early Childhood Environment Rating Scale (ECERS-R; Harms, Clifford, and Cryer 2005). Indicators from the CLASS use a 7-point Likert scale to measure teacher sensitivity and behavior management. Teacher sensitivity measures how responsive teachers are to children's academic and emotional needs. Teacher behavior management is an indicator of how well teachers monitor, prevent, and redirect children's behaviors in class. Based on 43 items, the ECERS-R is a widely used research tool that measures early childhood classroom quality across a wide range of constructs. In addition, the number of children observed in each classroom is used to control for the potential confounding effects of differences in class size. So too, a variable that measures the number of adults observed in each classroom is used as a control for differences in child-to-staff ratios.

Analytic Strategy

Propensity score matching is increasingly used to address selection bias in evaluating the effects of early childhood intervention programs (see research and reviews by Shonkoff and Phillips 2000; Hill et al. 2002, 2003, 2005; Lochman et al. 2006; Schneider et al. 2007; Guo and Fraser 2009; Zhai et al., forthcoming). A conventional propensity score matching approach uses observed pretreatment covariates to estimate the probability (i.e., the propensity score) that an individual will be assigned to the treatment group. The analysis then matches each member in the treatment group with the one or more control group members whose propensity scores are closest. There are several matching methods (e.g., nearest neighbor, radius, kernel, and Mahalanobis methods). If the analysis assumes that the predictive covariates are the only confounding variables, it can conceptualize these matched individuals with similar propensity scores as if they are randomly assigned to the treatment

or control group in an experiment (Rosenbaum and Rubin 1983, 1985; Heckman, Ichimura, and Todd 1997; Dehejia and Wahba 1999, 2002; Hill et al. 2002; Gibson 2003; Schneider et al. 2007).

This study uses a principal score matching method to match children in the control group with those who received high- or low-dosage treatment from the multifaceted CSRP intervention. Principal score matching is a method derived from propensity score matching. It builds on recent methodological innovations in principal stratification and subgroup analysis in the context of randomized experiments (Frangakis and Rubin 2002; Hill et al. 2002, 2003; Barnard et al. 2003; Gibson 2003; Peck 2003). Because two of the CSRP's three components (i.e., teacher training and MHC class visits) are provided at the class level, the strategy employed to analyze the class-level data differs from the strategy used to analyze data on the mental health consultation services provided individually to a small number of children.

Specifically, the analyses of the two class-level components of the CSRP intervention (i.e., teacher training and MHC class visits) include four stages. In the first stage, the study estimates the propensity of each treatment group classroom, j , to actually receive high- or low-dosage CSRP treatment (D). This first stage is conducted with the logit model specified in equation (1):

$$\text{logit}(D_j) = \beta_0 + \beta_1 C_j, \quad (1)$$

where C_j represents the pretreatment teacher and class characteristics in classroom j that possibly influence the class's propensity to receive high- or low-dosage CSRP treatment. These variables were collected prior to the CSRP intervention. The characteristics include teachers' personal stressors, work-related stressors, sensitivity, and behavior management; overall classroom quality (measured by ECERS-R scores); class size; and the number of adults in the classroom.

Using the coefficients obtained from equation (1), the study then estimates the dosage propensity scores for classrooms in the control group. These dosage propensity scores are referred to as principal scores, because they are used to stratify the population into mutually exclusive subgroups (or principal strata), which are based on pretreatment variables (Frangakis and Rubin 2002; Hill et al. 2003). Propensity scores differ from principal scores in the ways that they are estimated. In general, conventional propensity score matching involves an observed binary variable that indicates whether participants were in either the treatment or control group. Thus, the propensity scores estimated by a logit, logistic, or probit regression refer to the probabilities that all participants in the sample will receive the treatment of a program. In principal score matching, by contrast, control group participants' memberships in principal strata or subgroups (i.e., receiving high- or low-dosage services) cannot be observed directly (Frangakis and Rubin 2002; Gibson 2003; Peck 2003). Therefore, the analyses first estimate the treatment group's respective propensities to receive high- and low-dosage treatment (Hill et al. 2003). The resulting parameters are then applied to participants in the control group in order to estimate the respective probabilities that they

would receive high- and low-dosage treatment if they were assigned to the treatment group. These estimates are based on control group members' observed pretreatment characteristics, C_j (Hill et al. 2002, 2003; Gibson 2003; Peck 2003). Principal score matching is therefore appropriate for adjusting for such posttreatment variables as program dosage (Frangakis and Rubin 2002; Hill et al. 2003).

In the second stage, the principal scores estimated in the first stage are used to match classrooms in the treatment group with those control group classrooms that have the closest principal scores. The CSRP's RCT design ensures that classrooms in the treatment and control groups generally have similar characteristics, and these similarities make it possible to find control-treatment matches for classrooms with differential dosage levels (Hill et al. 2002, 2003; Peck 2003). The procedure uses a one-to-one nearest neighbor matching method. The analysis assumes that the predictors in equation (1) are the only confounding variables. Each pair of matched classrooms has similar principal scores, and thus the two matched classes in each pair are comparable in terms of the likelihood of receiving high- or low-dosage treatment (under the hypothetical condition that both were assigned to the treatment group).

Once the classrooms are matched, one intuitively appealing approach would be to continue the analyses at the class level and to estimate the dosage effects of treatment by comparing the average scores of children in pairs of matched classrooms. However, this approach faces additional challenges. Such an approach does not account for the variation in class-level treatment effects across individual children. Previous research finds that the CSRP treatment effects are moderated by child characteristics (e.g., gender, race or ethnicity, and poverty-related risk; Raver et al. 2009). In addition, comparing outcomes of matched classroom pairs is challenging because the study has a small sample of classrooms (i.e., 35 classrooms in total), and the numbers of classrooms in the high- and low-dosage treatment groups are also quite small (e.g., 10 classes in the low-dosage teacher training group and seven in the high-dosage group). These sizes offer little statistical power to detect effects (Raudenbush and Bryk 2002). Finally, comparing outcomes of matched classes is challenging because the focus of this study falls on the effects of CSRP intervention dosage on child school readiness outcomes. The intraclass correlations from unconditional three-level models (Lee 2000; Guo 2005; Trouilloud et al. 2006; Hedges and Hedberg 2007) suggest that the vast majority of variance in child outcomes is attributable to child-level heterogeneity (i.e., 73–78 percent in teacher-reported behavior problems and 83–95 percent in observer-rated social-emotional and cognitive scales). Therefore, child characteristics are likely to play a substantial role in children's school readiness. They also are likely to affect teachers' attendance at training sessions and MHCs' class visits. Matched models should therefore include child characteristics (Rosenbaum and Rubin 1983, 1985; Dehejia and Wahba 1999, 2002; Hill et al. 2002).

In the third stage, individual children in the matched classrooms are matched on their pretreatment characteristics. This step attempts to further account for child-level heterogeneity and is represented in equation (2):

$$\text{logit}(T_{ijk}) = \beta_0 + \beta_{ijk} X_{ijk} + \phi_k, \quad (2)$$

where T_{ijk} stands for the treatment status of child i in classroom j of matched classroom pair k (i.e., $T = 1$ as treatment, $T = 0$ as control), and X_{ijk} represents the child's pretreatment characteristics. These characteristics include gender, race and ethnicity (i.e., non-Hispanic black or not), poverty-related family risks, whether the child is from a single-parent family, whether his or her parents speak Spanish, and scores on pretreatment behavioral, social-emotional, and cognitive instruments. The fixed effect, ϕ_k , of matched classroom pair k represents the matching of children in the treatment group with control group counterparts who had similar pretreatment characteristics (conducted within the matched classrooms during the second stage). This procedure uses a one-to-one nearest neighbor matching method with replacement. Matching with replacement can minimize biases in estimates, because it allows each treatment unit to be matched with the nearest control unit, the latter of which can be matched again (i.e., matching with replacement) if it is the best match for other treatment units. Thus, the method produces higher match quality and is less sensitive to the order of units than matching without replacement (Abadie and Imbens 2002, 2006; Dehejia and Wahba 2002; Gibson 2003; Guo and Fraser 2009). Among children in the control group who had the same principal scores, one is chosen randomly for matching, and any nonmatched controls are discarded from further analysis (i.e., less than 3 percent in this study; Dehejia and Wahba 1999, 2002; Gibson 2003).

In the fourth stage, the dosage effects of the CSRP are estimated by calculating the regression-adjusted differences in the outcomes of matched children. Regression adjustment is increasingly used to estimate treatment effects in randomized experiments as well as in principal and propensity score matching methods. If one includes covariates in regression models after matching, one can reduce some bias from the matching by adjusting for the remaining differences in covariates among matched children. This approach also allows the investigator to take into account the effects of covariates on outcomes rather than to attribute differences in children's outcomes only to differences in their dosage levels of program participation (Abadie and Imbens 2002, 2006). As a result, adjusting for covariates after matching or randomized assignment can reduce potential bias and increase the chance of detecting statistically significant treatment effects (Rubin and Thomas 2000; Hill et al. 2002; Gibson 2003; Hill et al. 2003; Puma et al. 2005; Zhai et al., forthcoming). Regression adjustment after matching also reduces the chances of misspecification in the models because adjustments are made for relatively small differences in the covariates (Lu et al. 2001; Abadie and Imbens 2002, 2006). Therefore, ordinary least squares regressions are used to estimate the dosage effects of the CSRP in the sample of matched children. These estimates, which control for the pretreatment covariates of children, teachers, and classrooms, are presented in equation (3):

$$O_{ijk} = \beta_0 + \beta_1 D_{ijk} + \beta_2 X_{ijk} + \beta_3 C_{jk} + \phi_k + \xi_{ijk}, \quad (3)$$

where O_{ijk} represents the outcomes of child i in classroom j of matched classroom pair k , D_{ijk} stands for a binary variable of dosage measure (1 = high or low dosage; 0 = matched control), X_{ijk} denotes child characteristics, C_{jk} represents the covariates of teacher and class in classroom j , ϕ_k is a fixed effect of matched classroom pair k (showing that the comparison of outcome difference is among matched children who were in the same matched classrooms from the second stage), and ξ_{ijk} is a random error term. Huber-White robust standard errors are adopted to account for the cluster feature of children nested in classrooms. Because the matching method employs replacement in the third stage, the analysis also applies weights to equation (3); weights are calculated as the number of times that matched control units are used (Dehejia and Wahba 1999, 2002; Hill, Reiter, and Zanutto 2004; Zhai et al., forthcoming).

In contrast to the approach employed in the four-stage dosage analyses of the CSRP's two class-level components (i.e., teacher training and MHC class visits), a slightly different strategy is adopted for the dosage analyses of the child-level CSRP component (i.e., mental health consultation services for individual children). Because the individual MHC services are provided at the child level for a small number of children in all of the treatment group classrooms, principal score matching is conducted in three stages. In the first stage, equation (1) is used to estimate the propensities of children in the treatment group to receive individual MHC services. These propensities are estimated from children's pretreatment characteristics as well as from teacher and classroom covariates. Parameters obtained from equation (1) are then applied to children in the control group in order to estimate the probability that each receives individual mental health consultation services if he or she is assigned to the treatment group. In the second stage, the principal scores from the first stage are employed to match children who receive individual mental health consultation services to those in the control group who have similar principal scores. This step uses a one-to-one, nearest neighbor matching method with replacement. To account for Head Start site-level heterogeneity, children are matched within the same matched pairs of Head Start sites that the CSRP originally designed. Finally, in the third stage, regression-adjusted differences are employed (as specified in eq. [3]) to estimate the effects of individual mental health consultation services. In analyzing these effects on children's school readiness outcomes, individual mental health consultation services are treated as an extra dose of the CSRP intervention on children's school readiness outcomes.

The processes of estimating principal scores and matching are performed separately for low- and high-dosage treatment of two CSRP components (i.e., teacher training and MHC class visits) as well as for the extra dose of whether children received individual mental health consultation services. In a conventional propensity score matching approach, each individual person is associated with one propensity score that represents his or her probability of being treated. By contrast, the dosage analyses of the three CSRP components estimate principal scores for each individual person. Each individual's principal scores reflect his or her estimated propensities for specific dosage levels of treatment in each of the three CSRP components. This approach enables the analysis to estimate separately the respective dosage effects of individual components (Peck 2003).

Results

Descriptive Statistics

Table 1 presents the descriptive statistics on covariates by treatment condition and matching status. It also reports the significance levels from *t*-tests of the full control (column *a*: control group before matching) sample's differences from the treated samples (columns *b*: low- and high-dosage teacher training, low- and high-dosage MHC class visits, individual mental health consultation services). Significance levels for these tests are indicated in columns *b*. The table also presents the mean differences between the treated columns (columns *b*) and the matched columns (columns *c*: control groups after matching). The significance levels for these tests are indicated in columns *c*.

As the results in table 1 suggest, children in the low-dosage teacher training sample are more likely to be boys (54 percent) than are children in the full control group before matching (44 percent). The estimates also suggest that children in this low-dosage sample are more likely to be non-Hispanic black (i.e., 81 vs. 65 percent in the full control group) and to live with single parents (75 vs. 67 percent in the full control group). They are estimated to have more poverty-related risks (1.26 vs. .99 in the full control group) and higher pretreatment scores on early math skills. They are estimated to be less likely to have parents who speak Spanish (i.e., 17 vs. 25 percent in the full control group). The estimates further suggest that the classrooms in the low-dosage teacher training group tend to have teachers with more personal stressors (3.08) than do classrooms with teachers in the control group (2.45). So, too, teachers in the low-dosage classrooms are estimated to have lower behavior management skills (4.21 vs. 5.18) and less sensitivity (4.26 vs. 5.11); overall, the classrooms are of lower quality (4.08 vs. 4.97) than the control group classrooms.

By contrast, the estimates suggest that children in the high-dosage teacher training group are more likely than those in the full control group to be in some racial or ethnic category other than non-Hispanic black. Children in the high-dosage group also are estimated to have higher pretreatment scores on BPI internalizing and externalizing scales as well as on the PPVT. The results suggest that the high-dosage teacher training classrooms tend to have more adults than those in the full control group.

Compared to classrooms in the full control group, the treatment group classrooms that received low-dosage MHC class visits are estimated to have more boys and non-Hispanic black children. The estimates suggest that children in those low-dosage classrooms have more poverty-related risks and higher pretreatment scores on the BPI externalizing scale, the PPVT, and the measure of early math skills. Classrooms that received low-dosage MHC class visits also are estimated to have teachers with more personal stressors and lower behavior management skills than do the teachers in the control group classrooms.

Children in classrooms that received high-dosage MHC class visits are found to have higher pretreatment scores on the BPI internalizing and externalizing scales than do children in the full control group classrooms. Their scores on the attention and impulse control subscale are also found to be higher. The overall quality of classrooms that received high-dosage MHC class visits is estimated to be lower than that of control group classrooms, but classrooms

with high doses of MHC visits are estimated to have more adults. Finally, compared with children in the full control group, treatment group children who received individual mental health consultation services are estimated to be more likely to be boys; they are found to have more poverty-related risks and higher pretreatment scores on both BPI scales. In all treatment group classrooms, the number of children who received individual mental health consultation services is small. The estimates suggest that the teacher and classroom characteristics of children who received these individual services do not differ to a statistically significant degree from the teacher and classroom characteristics of children in the full control group. This actually is an overall comparison between the full treatment and the full control groups, since all treatment-assigned classes have children receiving individual mental health consultation services.

The results in table 1 also suggest that principal score matching leaves few statistically significant differences between matched control (cols. *c*) and treated (cols. *b*) groups. After matching, marginally significant mean differences (at $p < .10$) are found for two variables; non-Hispanic black children represent 39 percent of the high-dosage teacher training sample but only 51 percent of the matched control group. Results for the BPI internalizing scale suggest that children who received individual mental health consultation services have higher scores than their counterparts in the matched control groups (2.89 in the treated group vs. 2.19 in the matched control group). Nevertheless, the balance in these two variables improves considerably after matching. Matching is estimated to reduce the treatment-control difference by 22 percent for the variable that captures whether children are non-Hispanic black; it is found to decrease the difference by 10 percent for the variable that assesses the BPI internalizing scale score. In addition, treated and controlled classrooms do not show statistically significant differences in many covariates before matching. This is probably due to the low power of the small sample sizes of class-level data in detecting differences at conventional levels of statistical significance. The balance in most class-level covariates increases after matching. Therefore, the principal score matching approach employed in this study is generally able to identify comparable control groups for the low- and high-dosage treatment groups. The analyses within the matched samples might be able to reduce the selection biases on the observed covariates in estimates of the dosage effects of the CSRP intervention.

Dosage Effects of the CSRP Intervention

Table 2 presents a summary of the results on estimated CSRP dosage effects. It combines the estimates from the analyses of five data sets that are generated by multiple imputation. In terms of magnitudes and levels of statistical significance, the results are quite consistent and robust across the five data sets. For purposes of comparison, previous findings on CSRP's ITT effects (Raver et al. 2009, forthcoming) are presented in the second column. To account for the hierarchical structure of the CSRP data, the ITT effects are estimated using hierarchical linear modeling (Raudenbush and Bryk 2002). As the discussion mentions, children in these data are nested in classrooms and Head Start sites. In table 2, columns to the right of the ITT estimates present coefficients for low- and high-dosage teacher training, low- and high-dosage MHC class visits, and children who received individual mental health consultation services. To facilitate the comparison of the point estimates from different sets

of analyses in table 2, the corresponding effect sizes of these estimates are presented in the discussion below (without being shown in table 2). An effect size (ES) is measured in units of its SD (i.e., the point estimate divided by SD), which also makes it possible to compare the estimated effects in this study with those in other studies.

Table 2 suggests that, as previously hypothesized, several of the CSRP effects on school readiness are larger than the estimated ITT effects. Specifically, all of the estimated effects of high-dosage teacher training are greater than the corresponding effects estimated for ITT, and the effects of MHC class visits also are generally larger. The effects of low-dosage training levels are found to be smaller than estimated ITT effects or are not statistically significantly different from the ITT estimates. Moreover, as the discussion mentions, children who received individual mental health consultation services, which are an extra dose of the CSRP intervention in addition to teacher training and MHC class visits, are matched to their counterparts in the control group who had similar pretreatment characteristics. As shown in the last column of table 2, the estimated effects of individual mental health consultation services on school readiness measures are statistically significant.

In particular, the results from ITT analyses (table 2) suggest that most of the effects are statistically significant. Specifically, the ITT estimates suggest that the CSRP intervention is negatively and statistically significantly associated with scores on the two BPI scales (-1.81 on internalizing, corresponding to an ES of $-.91$ SDs; -2.92 on externalizing, ES = $-.64$ SDs). The ITT estimates also suggest that the CSRP intervention is positively and statistically significantly associated with scores on the attention and impulse control subscale (.20; ES = .37 SDs), the PPVT (1.46; ES = .33 SDs), and the early math skills measure (2.21; ES = .53 SDs). Table 2 further suggests that low-dosage teacher training, which accounts for 30–60 percent of the total training hours provided by the CSRP, has smaller estimated effects on children's BPI internalizing (-1.36 ; ES = $-.69$ SDs) and externalizing (-2.27 ; ES = $-.50$ SD) behavior problems than those identified in the ITT estimates, but the training effects are found to be statistically significant. Low-dosage training is estimated to be positively and statistically significantly associated with scores on PPVT (1.22; ES = .27 SDs) and early math skills (1.75; ES = .42 SDs). Both of those estimates are smaller than the corresponding ITT results. The coefficients for the attention and impulse control subscale and for positive emotion, still smaller than those in the ITT results, are estimated to be statistically nonsignificant.

In contrast, high-dosage teacher training, which accounts for 87–100 percent of the CSRP's total training hours, is found to have larger effects than those estimated for the ITT results on all six of the outcome variables. High-dosage training is estimated to be negatively and statistically significantly associated with scores on both BPI scales (-1.99 on internalizing, ES = -1.01 SDs; -4.06 on externalizing, ES = $-.89$ SDs). High-dosage training is estimated to be positively and statistically significantly associated with scores on the attention and impulse control subscale (.49; ES = .91 SDs), the PPVT (3.70; ES = .83 SDs), early math skills (3.74; ES = .90 SDs), and children's positive emotion (.40; ES = .76 SDs); the intervention is not statistically significantly associated with the positive emotion subscale score in ITT estimates.

Similarly, the results in table 2 suggest that the effects of low-dosage MHC class visits, which represent a level of exposure that falls below the average hours of MHC visits received in the treatment group, are smaller than those identified in ITT estimates on scores for both BPI scales, the attention and impulse control subscale, and early math skills. Low-dosage MHC visits are estimated to be negatively and statistically significantly associated with both internalizing (−1.12) and externalizing (−2.10) scales. The MHC visits are estimated to be positively and marginally associated with PPVT scores.

High-dosage MHC class visits represent a level of exposure that exceeds the average hours received by the treatment group. On all indicators of school readiness, high-dosage MHC visits are found to have larger effects than those identified in ITT estimates, and all of the coefficients are statistically significant.

Furthermore, the results in the last column of table 2 show that the individual mental health consultation services are estimated to have statistically significant effects on all measures of children's school readiness. Children in treatment-assigned classes who received individual mental health consultation services are estimated to have lower BPI scores and higher social-emotional and cognitive scores than their matched counterparts in the control group who have similar pretreatment characteristics.

The full regression-adjusted results from this study are presented in appendix table A1. These results are combined from the analyses of low- and high-dosage teacher training in the five imputed data sets. Although the magnitudes of associations and levels of statistical significance are estimated to be consistent and robust across the data sets, the standard errors are large, and most of the correlations are estimated to lose statistical significance when the results are combined. Analyses of the two CSRP components not shown in table A1 (MHC class visits and mental health consultation services for individual children) are found to produce very similar results. In addition, the authors also conducted dosage analyses that use the original data but do not employ imputation. The analyses identify patterns of dosage effects that are similar to the ones reported in this study, but the alternate analyses estimate smaller effects. Because missing data deprive these alternate analyses of statistical power, those results are also less likely to reach statistical significance.

Variations in Dosage Effects across Outcomes and CSRP Components

In addition, the results in table 2 also show the variations of dosage effects across school readiness measures and the CSRP intervention components. These estimates suggest that high-dosage teacher training is associated with larger reductions in children's externalizing behavior problems than are found in the ITT estimates. The effects of high-dosage training on social-emotional skills (scores on subscales for attention and impulse control and for positive emotion) and cognitive development (scores on the PPVT and early math skills) are estimated to be larger than those derived from ITT analyses. The estimates further suggest that low-dosage teacher training is not statistically significantly associated with either measure of children's social-emotional skills. Although statistically significant, the estimated effects of low-dosage training on internalizing behavior problems, externalizing behavior problems, PPVT scores, and early math skills are smaller than those derived from ITT analyses.

Similarly, low-dosage MHC class visits are estimated to predict smaller gains in children's school readiness outcomes (the association between low-dosage visits and the positive emotion score is found to be statistically nonsignificant; such visits are only estimated to be marginally associated with PPVT scores) than those derived from the ITT analyses or found in the high-dosage treatment estimates. In contrast, high-dosage MHC class visits are estimated to be statistically significantly related to all measured outcomes. The effects on reducing children's behavior problems, social-emotional skills, and cognitive development are all estimated to be larger than the effects derived from ITT estimates.

Moreover, results suggest that the individual mental health consultation services are especially efficient in promoting these children's social-emotional and PPVT skills. The effects of individual mental health consultation services on attention and impulse control, positive emotion, and PPVT scores are estimated to be larger than those derived from ITT estimates. The effects of individual mental health consultation services for children on the other measures of school readiness are estimated to be slightly smaller than those found in ITT estimates, but receipt of individual services is estimated to be statistically significantly associated with all of those outcomes.

The results in table 2 also permit comparisons of the effects estimated for teacher training with those estimated for MHC class visits. Compared with the effects estimated for low-dosage MHC visits, those estimated for low-dosage teacher training are larger on all measured outcomes except the score for the attention and impulse control subscale. So, too, the effects of high-dosage teacher training are estimated to be larger on all measured outcomes than those for high-dosage MHC class visits. The estimated effect sizes of individual mental health consultation services for children tend to fall between those of the low- and high-dosage levels of the other two components (i.e., teacher training and MHC class visits).

Conclusion and Discussion

Variations in compliance with social interventions, and in such interventions' effects on program outcomes, increasingly capture attention in the fields of policy evaluation and prevention science. Yet these important and policy-relevant empirical questions remain relatively understudied. This is due in part to the ways in which questions of dosage are complicated by the inferential challenges posed by individual differences in participants' propensity to take up or enroll in services. Using a principal score matching method to address the issue of selection bias, the current study examines the dosage effects of the CSRP, a classroom-based intervention conducted in Head Start settings located within seven disadvantaged Chicago neighborhoods. The analyses suggest several different solutions to the challenges of estimating the role of dosage across multiple program components of service delivery, within a cluster-randomized early intervention program, and at both the classroom and the child levels of analysis.

Using a principal score matching approach, the study finds that high-dosage levels of teacher training (i.e., 87–100 percent of the CSRP's total training hours) and high-dosage MHC class visits (i.e., above the average hours in the treatment group) are estimated to have larger

effects on children's school readiness than those derived from ITT analyses. In contrast, low-dosage levels of treatment (i.e., 30–60 percent of the total training hours for teacher training and below the average hours for MHC class visits) are predicted to have program effects that are smaller than those derived from ITT estimates or that are statistically nonsignificant. Moreover, individual mental health consultation services are estimated to be statistically significantly associated with the measured indicators of school readiness among children identified in a pretreatment assessment as having high levels of emotional and behavioral problems. In addition, dosage effects are found to vary across school readiness measures and the CSRP intervention components.

The findings in this study should be interpreted with caution. The CSRP intervention was conducted among a small sample of children who attended Head Start programs located in seven very disadvantaged neighborhoods in Chicago. The majority of the CSRP participants are reported to be either non-Hispanic black (66 percent) or Hispanic (26 percent). The sample thus did not share the ethnic composition of Head Start as a whole. In 2005 (the same time period as this evaluation), approximately 35 percent of enrolled Head Start participants were white, 31 percent were black, and 33 percent were Hispanic (Office of Head Start 2006). Moreover, research suggests that, across states and localities, there is considerable variation in preschool programs (Head Start and prekindergarten), the characteristics of enrolled children, and the effects of these programs on children's developmental outcomes (Gormley 2007; Rigby, Ryan, and Brooks-Gunn 2007; Wong et al. 2008). In addition, the design of the CSRP intervention, which is conducted in Head Start program settings, might influence its treatment effects in ways that differ from the influences exerted in other preschool settings. Therefore, the study's findings on the program effects of the CSRP intervention, including the estimates of ITT and dosage effects, should be understood to represent a specific context: a single efficacy trial in 18 Head Start programs in Chicago. As such, the results should not be generalized more broadly. To check the robustness of the findings in this study and to provide more generalizable results, future research should employ samples that are more demographically and geographically representative of the national population of preschool-aged, economically disadvantaged children. Such research should occur within the context of a large-scale intervention.

Results also suggest that the dosage measures for the three components of the CSRP intervention are relatively independent from each other. For this reason, the dosage effects of individual components are estimated under the assumption that the participants also received the services of other components, the dosage levels of which were distributed equally or randomly. For example, in comparing the outcomes of children in the low-dosage teacher training group with those of children in the control group who had similar background characteristics, one should keep in mind that children in the low-dosage teacher training group received MHC class visits as well, and some of them also received individual mental health consultation services. As the findings from ITT and individual dosage analyses suggest, all three components of the CSRP intervention could contribute to the promotion of children's school readiness. The estimates of the respective roles of the study's individual CSRP components might be larger than the estimates obtained if these individual components of services were provided separately as single packages of intervention to different groups of children. Nevertheless, due to the small sample size and the multifaceted

design of the CSRP intervention, it was not possible to detect the pure effects of different dosage levels in individual program components. If fiscally and practically feasible, an important next empirical step would be for researchers to conduct efficacy trials in which individual components are unbundled and implemented separately for different groups of children. Such a design would help identify the components' ITT as well as dosage effects.

As a derivative of the propensity score matching approach, a principal score matching method is subject to the assumption of ignorable treatment or selection on observables. This approach requires observation of all confounding covariates related to treatment status (Rosenbaum and Rubin 1983; Dehejia and Wahba 1999, 2002; Joffe and Rosenbaum 1999; Hill et al. 2002, 2003, 2005; Gibson 2003). If any important covariates are omitted in the predictive models (e.g., eq. [1]), then group members could be mismatched, and the estimates of treatment effects could be biased. For example, teachers' motivation might represent an important omitted variable that could drive some teachers to be more highly engaged in training and MHC services (i.e., to be more likely to receive high-dosage treatment) than are other, less motivated teachers. So too, motivation might prompt some to do a better job of fostering positive outcomes among the children enrolled in their classrooms. Similar to many other early childhood studies, the CSRP did not collect data on teachers' motivation. The current analyses attempt to include a number of variables that are likely associated with teachers' motivation. For example, the analyses include variables on perception of job demand, job control, resources, and personal stressors, as well as measures of teachers' education and work-related stressors that reflect self-efficacy (Ghaith and Yaghi 1997; Cross and Wyman 2006; Baker et al. 2010). Despite the inclusion of those variables, the study cannot rule out the inferential threat posed by unobserved variables in using principal score matching techniques. As with most approaches to dosage analyses, the study's estimates of CSRP program dosage effects rely on a set of strong assumptions. As a result, the estimates of relative effects of high, low, and no dosage may be inflated. Therefore, these findings should be interpreted with great caution.

In addition, this study estimates principal scores by running the predictive model in the treatment group first and then using the obtained coefficients to fit it to the control group. The issue of overfitting poses concerns because the predictive model would provide a better fit for a subsample (i.e., the treatment group) than for the rest of the sample (i.e., the control group). This could result in possible biases in prediction (Gibson 2003; Peck 2003). One way to address this issue is to create an external sample to estimate the predictive model and then to fit the model to both the remaining treatment and control groups (Gibson 2003; Peck 2003). However, the CSRP is similar to many other randomized social experiments (e.g., Hill et al. 2002, 2003; Gibson 2003) in that its sample is not large enough to allow the creation of an external subsample for prediction and then to exclude that external subsample from further matching and dosage analyses. As the discussion indicates above, large-scale and carefully designed interventions may eliminate such potential biases and provide more robust findings.

Given these limitations, what are the policy implications of this study's findings? In the authors' view, ITT estimates help policy professionals and practitioners to set empirically conservative benchmarks for gains that children can reasonably be expected to make under

real-world conditions of program participation. In contrast, principal score matching estimates help to identify upper bound estimates of what might be achieved through a best-case scenario of comprehensive intervention. Specifically, principal score matching techniques provide policy professionals with an estimate of the effect sizes that they might expect to find if programs were more successful at providing a full rather than partial dosage of the intervention to a larger fraction of the enrolled sample. For example, the current study provides evidence that high-dosage classroom-based intervention in Head Start settings can result in estimates of larger effects on children's school readiness than those obtained from ITT estimates. The results also suggest that low-dosage levels of participation have smaller effects than those obtained with ITT estimates or that they have no statistically significant program effects. These findings are consistent with results from prior studies that use similar methodologies and target behavioral and other developmental outcomes of children at risk (e.g., Hill et al. 2003; Lochman et al. 2006). A clear implication of these findings is that policy makers and program staff might be motivated to increase program participation levels, particularly in classroom-based interventions for at-risk children and their teachers.

Conversely, the results on low-dosage teacher training and MHC class visits provide policy makers with lower bound estimates of the CSRP intervention's effects, and these estimates are even more conservative than the ITT estimates. Although policy makers would need to keep in mind the caveats discussed above, these findings provide encouraging evidence should they choose to support the implementation of the CSRP model in other types of early childhood programs and in other locales. It is notable that even these lower-bound estimates are mostly moderate in magnitude. Moreover, comparisons of the lower and upper bound estimates suggest what types of program components policy makers might especially target with limited resources. For instance, high doses of teacher training are estimated to yield large improvements in children's attention and positive emotions, whereas low doses of teacher training are found to garner no improvement in these outcomes. On other outcomes, the difference between the effects of high-dosage teacher training and those of low-dosage teacher training are smaller. Policy makers interested in fostering the development of children's self-regulatory competence may find it helpful to know that supporting teachers' participation in this type of professional development is particularly important.

These findings, combined with empirical analyses of the obstacles that may limit program participation, shift attention from the most pressing policy question of whether programs like Head Start work (e.g., Zigler and Styfco 2004; Barnett 2007; Gormley 2007) to a nuanced policy consideration of the ways in which programs may work better for some participants and less well for others. These analyses provide helpful guideposts in considering alternative scenarios of service delivery. They may help researchers to consider the challenges and the benefits of using multifaceted classroom-based services to target low-income children's chances of school success. The authors hope that these findings spark new debate and discussion on ways that policy makers, prevention scientists, and practitioners can take important next steps in improving the lives of low-income children living in communities of concentrated economic disadvantage.

Acknowledgments

The project described was supported by award number R01HD046160 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Eunice Kennedy Shriver National Institute of Child Health and Human Development or the National Institutes of Health. The Chicago School Readiness Project is not associated with The Chicago School, which is a trademark of the Chicago School of Professional Psychology.

References

- Abadie, Alberto; Imbens, Guido W. Technical Working Paper no 283. National Bureau of Economic Research; Cambridge, MA: 2002. Simple and Bias-Corrected Matching Estimators for Average Treatment Effects.
- Abadie, Alberto; Imbens, Guido W. Large Sample Properties of Matching Estimators for Average Treatment Effects. *Econometrica*. 2006; 74(1):235–67.
- Aber, J Lawrence; Brown, Joshua L.; Jones, Stephanie M. Developmental Trajectories toward Violence in Middle Childhood: Course, Demographic Differences, and Response to School-Based Intervention. *Developmental Psychology*. 2003; 39(2):324–48. [PubMed: 12661889]
- Agodini, Roberto; Dynarski, Mark. Are Experiments the Only Option? A Look at Dropout Prevention Programs. *Review of Economics and Statistics*. 2004; 86(1):180–94.
- Alkon, Abbey; Ramler, Malia; MacLennan, Katharine. Evaluation of Mental Health Consultation in Child Care Centers. *Early Childhood Education Journal*. 2003; 31(2):91–99.
- Angold, Adrian; Jane Costello, E.; Burns, Barbara J.; Erkanli, Alaattin; Farmer, Elizabeth MZ. Effectiveness of Nonresidential Specialty Mental Health Services for Children and Adolescents in the 'Real World.'. *Journal of the American Academy of Child and Adolescent Psychiatry*. 2000; 39(2):154–60. [PubMed: 10673824]
- Angrist, Joshua D. Instrumental Variables Methods in Experimental Criminological Research: What, Why and How. *Journal of Experimental Criminology*. 2006; 2(1):23–44.
- Baker, Courtney N.; Kupersmidt, Janis B.; Voegler-Lee, Mary Ellen; Arnold, David H.; Willoughby, Michael T. Predicting Teacher Participation in a Classroom-Based, Integrated Preventive Intervention for Preschoolers. *Early Childhood Research Quarterly*. 2010; 25(3):270–83. [PubMed: 21103189]
- Barkley, Russell A.; Shelton, Terri L.; Crosswait, Cheryl; Moorehouse, Maureen; Fletcher, Kenneth; Barrett, Susan; Jenkins, Lucy; Metevia, Lori. Multi-Method Psycho-Educational Intervention for Preschool Children with Disruptive Behavior: Preliminary Results at Post-treatment. *Journal of Child Psychology and Psychiatry*. 2000; 41(3):319–32. [PubMed: 10784079]
- Barnard, John; Frangakis, Constantine E.; Hill, Jennifer L.; Rubin, Donald B. Principal Stratification Approach to Broken Randomized Experiments: A Case Study of School Choice Vouchers in New York City. *Journal of the American Statistical Association*. 2003; 98(462):299–311.
- Barnett, W Steven. Revving Up Head Start: Lessons from Recent Research. *Journal of Policy Analysis and Management*. 2007; 26(3):674–77.
- Bickman, Leonard; Andrade, Ana Regina; Warren Lambert, E. Dose Response in Child and Adolescent Mental Health Services. *Mental Health Services Research*. 2002; 4(2):57–70. [PubMed: 12090308]
- Bloom, Howard S. *Learning More from Social Experiments: Evolving Analytic Approaches*. New York: Russell Sage; 2005.
- Brennan, Eileen M.; Bradley, Jennifer R.; Allen, Mary Dallas; Perry, Deborah F. The Evidence Base for Mental Health Consultation in Early Childhood Settings: Research Synthesis Addressing Staff and Program Outcomes. *Early Education and Development*. 2008; 19(6):982–1022.
- Brooks-Gunn, Jeanne; Duncan, Greg J.; Lawrence Aber, J. *Neighborhood Poverty: Context and Consequences for Children*. New York: Russell Sage; 1997.
- Cathers-Schiffman, Teresa A.; Thompson, Marilyn S. Assessment of English-and Spanish-Speaking Students with the WISC-III and Leiter-R. *Journal of Psychoeducational Assessment*. 2007; 25(1): 41–52.

- Cross, Wendi; Wyman, Peter A. Training and Motivational Factors as Predictors of Job Satisfaction and Anticipated Job Retention among Implementers of a School-Based Prevention Program. *Journal of Primary Prevention*. 2006; 27(2):195–215. [PubMed: 16550457]
- Curbow, Barbara; Spratt, Kai; Ungaretti, Antoinette; McDonnell, Karen; Breckler, Steven. Development of the Child Care Worker Job Stress Inventory. *Early Childhood Research Quarterly*. 2000; 15(4):515–36.
- Dehejia, Rajeev H.; Wahba, Sadek. Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs. *Journal of the American Statistical Association*. 1999; 94(448): 1053–62.
- Dehejia, Rajeev H.; Wahba, Sadek. Propensity Score-Matching Methods for Nonexperimental Causal Studies. *Review of Economics and Statistics*. 2002; 84(1):151–61.
- Dunn, Lloyd M.; Dunn, Leota M. Peabody Picture Vocabulary Test, Third Edition (PPVT-III). Circle Pines, MN: American Guidance Service; 1997.
- Frangakis, Constantine E.; Rubin, Donald B. Principal Stratification in Causal Inference. *Biometrics*. 2002; 58(1):21–29. [PubMed: 11890317]
- Ghaith G, Yaghi H. Relationships among Experience, Teacher Efficacy, and Attitudes toward the Implementation of Instructional Innovation. *Teaching and Teacher Education*. 1997; 13(4):451–58.
- Gibson, Christina M. Privileging the Participant: The Importance of Sub-group Analysis in Social Welfare Evaluations. *American Journal of Evaluation*. 2003; 24(4):443–69.
- Gormley, William T, Jr. Early Childhood Care and Education: Lessons and Puzzles. *Journal of Policy Analysis and Management*. 2007; 26(3):633–71.
- Gottfredson, Gary D.; Jones, Elizabeth M.; Gore, Thomas W. Implementation and Evaluation of a Cognitive-Behavioral Intervention to Prevent Problem Behavior in a Disorganized School. *Prevention Science*. 2002; 3(1):43–56. [PubMed: 12002558]
- Green, Beth L.; Everhart, Maria; Gordon, Lyn; Gettman, Maria Garcia. Characteristics of Effective Mental Health Consultation in Early Childhood Settings: Multilevel Analysis of a National Survey. *Topics in Early Childhood Special Education*. 2006; 26(3):142–52.
- Guo, Shenyang. Analyzing Grouped Data with Hierarchical Linear Modeling. *Children and Youth Services Review*. 2005; 27(6):637–52.
- Guo, Shenyang; Fraser, Mark W. Propensity Score Analysis: Statistical Methods and Applications. Thousand Oaks, CA: Sage; 2009.
- Hammarberg, Annie; Hagekull, Berit. The Relation between Pre-school Teachers' Classroom Experiences and Their Perceived Control over Child Behaviour. *Early Child Development and Care*. 2002; 172(6):625–34.
- Harms, Thelma; Clifford, Richard M.; Cryer, Debby. Early Childhood Environment Rating Scale, Revised Edition (ECERS-R). New York: Teachers College Press; 2005.
- Heckman, James J.; Ichimura, Hidehiko; Todd, Petra E. Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme. *Review of Economic Studies*. 1997; 64(4):605–54.
- Hedges, Larry V.; Hedberg, EC. Intraclass Correlations for Planning Group Randomized Experiments in Rural Education. *Journal of Research in Rural Education*. 2007; 22(10)
- Hill, Jennifer L.; Brooks-Gunn, Jeanne; Waldfogel, Jane. Sustained Effects of High Participation in an Early Intervention for Low-Birth-Weight Premature Infants. *Developmental Psychology*. 2003; 39(4):730–44. [PubMed: 12859126]
- Hill, Jennifer L.; Reiter, Jerome P.; Zanutto, Elaine L. A Comparison of Experimental and Observational Data Analyses. In: Gelman, Andrew; Meng, Xiao-Li, editors. *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family*. West Sussex: Wiley; 2004. p. 49–60.
- Hill, Jennifer L.; Waldfogel, Jane; Brooks-Gunn, Jeanne. Differential Effects of High-Quality Child Care. *Journal of Policy Analysis and Management*. 2002; 21(4):601–27.
- Hill, Jennifer L.; Waldfogel, Jane; Brooks-Gunn, Jeanne; Han, Wen-Jui. Maternal Employment and Child Development: A Fresh Look Using Newer Methods. *Developmental Psychology*. 2005; 41(6):833–50. [PubMed: 16351331]

- Hooper, V Scott; Bell, Sherry Mee. Concurrent Validity of the Universal Non-verbal Intelligence Test and the Leiter International Performance Scale-Revised. *Psychology in the Schools*. 2006; 43(2): 143–48.
- Imbens, Guido W. The Role of the Propensity Score in Estimating Dose-Response Functions. *Biometrika*. 2000; 87(3):706–10.
- Joffe, Marshall M.; Rosenbaum, Paul R. Invited Commentary: Propensity Scores. *American Journal of Epidemiology*. 1999; 150(4):327–33. [PubMed: 10453808]
- La Paro, Karen M.; Pianta, Robert C.; Stuhlman, Megan. The Classroom Assessment Scoring System: Findings from the Prekindergarten Year. *Elementary School Journal*. 2004; 104(5):409–26.
- Lee, Valerie E. Using Hierarchical Linear Modeling to Study Social Contexts: The Case of School Effects. *Educational Psychologist*. 2000; 35(2):125–41.
- Li-Grining, Christine P.; Votruba-Drzal, Elizabeth; Bachman, Heather J.; Lindsay Chase-Lansdale, P. Are Certain Preschoolers at Risk in the Era of Welfare Reform? The Moderating Role of Children's Temperament. *Children and Youth Services Review*. 2006; 28(9):1102–23.
- Little, Roderick JA.; Rubin, Donald B. *Statistical Analysis with Missing Data*. New York: Wiley; 1987.
- Lochman, John E.; Boxmeyer, Caroline; Powell, Nicole; Roth, David L.; Windle, Michael. Masked Intervention Effects: Analytic Methods for Addressing Low Dosage of Intervention. *New Directions for Evaluation*. 2006; 110:19–32.
- Lochman, John E.; Wells, Karen C. Effectiveness of the Coping Power Program and of Classroom Intervention with Aggressive Children: Outcomes at a 1-Year Follow-Up. *Behavior Therapy*. 2003; 34(4):493–515.
- Love, John M.; Kisker, Ellen Eliason; Ross, Christine M.; Schochet, Peter Z.; Brooks-Gunn, Jeanne; Paulsell, Diane; Boller, Kimberly; Constantine, Jill; Vogel, Cheri; Fuligni, Allison Sidle; Brady-Smith, Christy. *Making a Difference in the Lives of Infants and Toddlers and Their Families: The Impacts of Early Head Start*. U.S. Department of Health and Human Services, Head Start Bureau; Washington, DC: 2002 Jun. Report
- Lu, Bo; Zanutto, Elaine; Hornik, Robert; Rosenbaum, Paul R. Matching with Doses in an Observational Study of a Media Campaign against Drug Abuse. *Journal of the American Statistical Association*. 2001; 96(456):1245–53. [PubMed: 25525284]
- Markowitz, Joy; Carlson, Elaine; Frey, William; Riley, Jarnee; Shimshak, Amy; Heinzen, Harriotte; Strohl, Jeff; Klein, Sheri; Lee, Hyunshik. *Preschoolers with Disabilities: Characteristics, Services, and Results; Wave 1 Overview Report from the Pre-elementary Education Longitudinal Study (PEELS)*. U.S. Department of Education, National Center for Special Education Research; Washington, DC: 2006. Report no. NCSE 2006–3003
- Office of Head Start. *Head Start Program Fact Sheet*. U.S. Department of Health and Human Services, Administration for Children and Families, Head Start Bureau; Washington, DC: 2006. <http://www.acf.hhs.gov/programs/ohs/about/fy2006.html>
- Peck, Laura R. Subgroup Analysis in Social Experiments: Measuring Program Impacts Based on Post-treatment Choice. *American Journal of Evaluation*. 2003; 24(2):157–87.
- Perry, Deborah F.; Clare Dunne, M.; McFadden, La Tanya; Campbell, Doreen. Reducing the Risk for Preschool Expulsion: Mental Health Consultation for Young Children with Challenging Behaviors. *Journal of Child and Family Studies*. 2008; 17(1):44–54.
- Puma, Michael; Bell, Stephen; Cook, Ronna; Heid, Camila; Lopez, Michael; Zill, Nicholas; Shapiro, Gary. *Head Start Impact Study: First Year Findings*. U.S. Department of Health and Human Services, Office of Planning, Research and Evaluation; Washington, DC: 2005 May. Report
- Raudenbush, Stephen W.; Bryk, Anthony S. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Thousand Oaks, CA: Sage; 2002.
- Raver, C Cybele. *Emotions Matter: Making the Case for the Role of Young Children's Emotional Development for Early School Readiness*. Social Policy Report (Journal of the Society for Research in Child Development). 2002; 16(4)
- Raver, C Cybele. Does Work Pay Psychologically as Well as Economically? The Role of Employment in Predicting Depressive Symptoms and Parenting among Low-Income Families. *Child Development*. 2003; 74(6):1720–36. [PubMed: 14669892]

- Raver, C Cybele; Garner, Pamela W.; Smith-Donald, Radiah. The Roles of Emotion Regulation and Emotion Knowledge for Children's Academic Readiness: Are the Links Causal?. In: Pianta, Robert C.; Cox, Martha J.; Snow, Kyle L., editors. *School Readiness and the Transition to Kindergarten in the Era of Accountability*. Baltimore: Brookes; 2007. p. 121-47.
- Raver, C Cybele; Jones, Stephanie M.; Li-Grining, Christine; Zhai, Fuhua; Bub, Kristen; Pressler, Emily. CSRP's Impact on Low-Income Preschoolers' Pre-academic Skills: Self-Regulation and Teacher-Student Relationships as Two Mediating Mechanisms. *Child Development*. Forthcoming.
- Raver, C Cybele; Jones, Stephanie M.; Li-Grining, Christine; Zhai, Fuhua; Metzger, Molly W.; Solomon, Bonnie. Targeting Children's Behavior Problems in Preschool Classrooms: A Cluster-Randomized Controlled Trial. *Journal of Consulting and Clinical Psychology*. 2009; 77(2):302-16. [PubMed: 19309189]
- Reid, M Jamila; Webster-Stratton, Carolyn; Hammond, Mary. Follow-Up of Children Who Received the Incredible Years Intervention for Oppositional-Defiant Disorder: Maintenance and Prediction of 2-Year Outcome. *Behavior Therapy*. 2003; 34(4):471-91.
- Rigby, Elizabeth; Ryan, Rebecca M.; Brooks-Gunn, Jeanne. Child Care Quality in Different State Policy Contexts. *Journal of Policy Analysis and Management*. 2007; 26(4):887-907.
- Roid, Gale H.; Miller, Lucy. *Leiter International Performance Scale-Revised (Leiter-R)*. Wood Dale, IL: Stoelting; 1997.
- Rosenbaum, Paul R.; Rubin, Donald B. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*. 1983; 70(1):41-55.
- Rosenbaum, Paul R.; Rubin, Donald B. Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. *American Statistician*. 1985; 39(1):33-38.
- Royston, Patrick. Multiple Imputation of Missing Values: Update of ICE. *Stata Journal*. 2005; 5(4): 527-36.
- Rubin, Donald B. Inference and Missing Data. *Biometrika*. 1976; 63(3):581-92.
- Rubin, Donald B. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley; 1987.
- Rubin, Donald B.; Thomas, Neal. Combining Propensity Score Matching with Additional Adjustments for Prognostic Covariates. *Journal of the American Statistical Association*. 2000; 95(450):573-85.
- Schaeffer, Cindy M.; Petras, Hanno; Ialongo, Nicholas; Masyn, Katherine E.; Hubbard, Scott; Poduska, Jeanne; Kellam, Sheppard. A Comparison of Girls' and Boys' Aggressive-Disruptive Behavior Trajectories across Elementary School: Prediction of Young Adult Antisocial Outcomes. *Journal of Consulting and Clinical Psychology*. 2006; 74(3):500-550. [PubMed: 16822107]
- Schafer, Joseph L. *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall; 1997.
- Schneider, Barbara; Carnoy, Martin; Kilpatrick, Jeremy; Schmidt, William H.; Shavelson, Richard J. *Think Tank White Paper*. American Educational Research Association; Washington, DC: 2007.
- Estimating Causal Effects Using Experimental and Observational Designs.
- Scott-Little, M Catherine; Holloway, Susan D. Child Care Providers' Reasoning about Misbehaviors: Relation to Classroom Control Strategies and Professional Training. *Early Childhood Research Quarterly*. 1992; 7(4):595-606.
- Shermoff, Elisa Steele; Kratochwill, Thomas R. Transporting an Evidence-Based Classroom Management Program for Preschoolers with Disruptive Behavior Problems to a School: An Analysis of Implementation, Outcomes, and Contextual Variables. *School Psychology Quarterly*. 2007; 22(3):449-72.
- Shonkoff, Jack P.; Phillips, Deborah A., editors. *From Neurons to Neighborhoods: The Science of Early Childhood Development*. Washington, DC: National Academy Press; 2000.
- Smith, Jeffrey A.; Todd, Petra E. Does Matching Overcome LaLonde's Critique of Non-experimental Estimators? *Journal of Econometrics*. 2005; 125(1-2):305-53.
- Smith-Donald, Radiah; Cybele Raver, C.; Hayes, Tiffany; Richardson, Breeze. Preliminary Construct and Concurrent Validity of the Preschool Self-Regulation Assessment (PSRA) for Field-Based Research. *Early Childhood Research Quarterly*. 2007; 22(2):173-87.
- Smolkowski, Keith; Biglan, Anthony; Barrera, Manuel; Taylor, Ted; Black, Carol; Blair, Jason. *Schools and Homes in Partnership (SHIP): Long-Term Effects of a Preventive Intervention*

- Focused on Social Behavior and Reading Skill in Early Elementary School. *Prevention Science*. 2005; 6(2):113–25. [PubMed: 15889626]
- Spencer, Michael S.; Fitch, Dale; Grogan-Kaylor, Andrew; McBeath, Bowen. The Equivalence of the Behavior Problem Index across U.S. Ethnic Groups. *Journal of Cross-Cultural Psychology*. 2005; 36(5):573–89.
- Tolan, Patrick; Gorman-Smith, Deborah; Henry, David. Supporting Families in a High-Risk Setting: Proximal Effects of the SAFE Children Preventive Intervention. *Journal of Consulting and Clinical Psychology*. 2004; 72(5):855–69. [PubMed: 15482043]
- Trouilloud, David; Sarrazin, Philippe; Bressoux, Pascal; Bois, Julien. Relation between Teachers' Early Expectations and Students' Later Perceived Competence in Physical Education Classes: Autonomy-Supportive Climate as a Moderator. *Journal of Educational Psychology*. 2006; 98(1): 75–86.
- Van Buuren S, Boshuizen HC, Knook DL. Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis. *Statistics in Medicine*. 1999; 18(6):681–94. [PubMed: 10204197]
- Wakschlag, Lauren S.; Leventhal, Bennett; Briggs-Gowan, Margaret; Danis, Barbara; Keenan, Kate; Hill, Carri; Egger, Helen; Cicchetti, Domenic; Carter, Alice. Defining the 'Disruptive' in Preschool Behavior: What Diagnostic Observation Can Teach Us. *Clinical Child and Family Psychology Review*. 2005; 8(3):183–201. [PubMed: 16151617]
- Waldfoegel, Jane. *What Children Need*. Cambridge, MA: Harvard University Press; 2006.
- Webster-Stratton, Carolyn; Reid, M Jamila; Hammond, Mary. Preventing Conduct Problems, Promoting Social Competence: A Parent and Teacher Training Partnership in Head Start. *Journal of Clinical Child Psychology*. 2001; 30(3):283–302. [PubMed: 11501247]
- Webster-Stratton, Carolyn; Jamila Reid, M.; Hammond, Mary. Treating Children with Early-Onset Conduct Problems: Intervention Outcomes for Parent, Child, and Teacher Training. *Journal of Clinical Child and Adolescent Psychology*. 2004; 33(1):105–24. [PubMed: 15028546]
- Webster-Stratton, Carolyn; Taylor, Ted. Nipping Early Risk Factors in the Bud: Preventing Substance Abuse, Delinquency, and Violence in Adolescence through Interventions Targeted at Young Children (0–8 Years). *Prevention Science*. 2001; 2(3):165–92. [PubMed: 11678292]
- Williford, Amanda P; Shelton, Terri L. Using Mental Health Consultation to Decrease Disruptive Behaviors in Preschoolers: Adapting an Empirically-Supported Intervention. *Journal of Child Psychology and Psychiatry*. 2008; 49(2):191–200. [PubMed: 18211278]
- Wong, Vivian C.; Cook, Thomas D.; Steven Barnett, W.; Jung, Kwanghee. An Effectiveness-Based Evaluation of Five State Pre-kindergarten Programs. *Journal of Policy Analysis and Management*. 2008; 27(1):122–54.
- Yeung, W Jean; Linver, Miriam R.; Brooks-Gunn, Jeanne. How Money Matters for Young Children's Development: Parental Investment and Family Processes. *Child Development*. 2002; 73(6):1861–79. [PubMed: 12487499]
- Zhai, Fuhua; Brooks-Gunn, Jeanne; Waldfoegel, Jane. Head Start and Urban Children's School Readiness: A Birth Cohort Study in 18 Cities. *Developmental Psychology*. Forthcoming.
- Zigler, Edward; Styfco, Sally J. Applying the Findings of Developmental Psychology to Improve Early Childhood Intervention. In: Feldman, Maurice A., editor. *Early Intervention: The Essential Readings*. Malden, MA: Blackwell; 2004. p. 54-72.
- Zill, Nicholas. *The Behavior Problems Index*. Washington, DC: Child Trends; 1990.
- Zill, Nicholas. *Early Math Skills Test*. Rockville, MD: Westat; 2003a.
- Zill, Nicholas. *Letter Naming Task*. Rockville, MD: Westat; 2003b.

Appendix. Missing Data and Multiple Imputation

As the main text indicates, 602 children participated in the CSRP. Missing data are common in empirical studies with human subjects; data on outcome measures and control variables are missing for a small proportion of children in this study. There is variation in the number of children for whom data are available on outcomes measured in May of the Head Start year. Specifically, the number ranges from a low of 498 children with data on the PPVT

scale to a high of 547 children with data on the two BPI scales. If separate complete case analyses were conducted for each outcome variable, the samples in these analyses would be different. This would hinder comparisons across analyses and be particularly problematic for studying trends over time (Hill et al. 2002). Therefore, the current analyses are limited to children for whom data are available on all outcome measures in May of the Head Start year. The resulting sample comprises 489 children.

Full data are available for most control variables. Specifically, complete data on teacher- and class-level covariates are available for all children. For small proportions of the sample, data are missing on poverty-related family risks (10 percent), on whether the participant is part of a single-parent family (5 percent), and on pretreatment scores (12 percent on teacher-reported BPI and 16–19 percent on observer-reported scales that include attention and impulse control, positive emotion, PPVT, and early math skills). As a result, complete case analyses on all outcome variables and covariates would reduce the sample to 372 children, as the sample would exclude a large portion (38 percent) of the original sample of 602 children. Such a dramatic loss would leave a sample with poor statistical power to detect treatment effects and to perform principal score matching (Little and Rubin 1987; Hill et al. 2002, 2004; Raudenbush and Bryk 2002).

To address the issue of missing data, multiple imputation uses multiple predictions for each missing value on certain variables. These predictions are based on other observed variables and are made to account for the uncertainty in imputed values (Rubin 1987; Schafer 1997; Hill et al. 2002, 2004; Guo and Fraser 2009). This study assumes that data are missing at random (Rubin 1976; Little and Rubin 1987; Van Buuren, Boshuizen, and Knook 1999; Hill et al. 2004). The assumption is plausible because the estimates suggest that CSRP children for whom data are missing do not differ from counterparts with complete data (Raver et al. 2009).

The ICE command in Stata statistical software (version 10) is used to conduct multiple imputation. This command implements multiple imputation by chained equations (Van Buuren et al. 1999; Royston 2005). All child-level variables and class-fixed effects are included. Using class-fixed effects models in multiple imputation allows the analyses to control for the unobserved heterogeneity across classrooms. In addition, a bootstrap method is adopted for creating imputed values. The method estimates regression coefficients in a bootstrap sample of the nonmissing observations and thus has the advantage of robustness (Van Buuren et al. 1999).

The analyses generate five sets of imputations for missing data and perform separate principal score matching with each data set (Little and Rubin 1987; Rubin 1987; Schafer 1997; Hill et al. 2004). As the discussion above details, data on one or more variables are missing for approximately 5–19 percent of children in this study. For five imputed data sets, the expected relative efficiency for recovering missing values ranges from 98.2 to 99.5 percent. This is based on equation (A1) (Rubin 1987):

$$RE = \left(1 + \frac{\gamma}{M}\right)^{-1/2}, \quad (\text{A1})$$

where γ is the fraction of missing information and M is a finite number of imputation ($M = 5$ in this study).

After obtaining the estimates of separate dosage effects from the analyses of five imputed data sets, the study uses their means as the final estimates of dosage effects. The standard errors are obtained using Donald Rubin's (1987) rules for combining multiple imputation. This procedure is shown in equation (A2):

$$\hat{s}_M = \left[M^{-1} \sum_m \hat{s}_m^2 + (1 + M^{-1})(M - 1)^{-1} \sum_m (\hat{b}_m - \bar{b})^2 \right]^{1/2}, \quad (\text{A2})$$

where $\hat{b}_m$ is the estimated coefficient of high- or low-dosage treatment; the standard error is $\hat{s}_m$ in sample m of M imputed samples ($M = 5$ in this study); and $\bar{b}$ is the mean of the coefficients estimated from the imputed samples. That mean is also the final estimate of dosage effect.

Table A1

Full Regression-Adjusted Results from the Analyses of Low- and High-Dosage Teacher Training

	Low Dosage							
	BPI-I	BPI-E	A-I	Pos Emo	PPVT	Math	BPI-I	BPI-E
Treatment	-1.36** (.46)	-2.27* (.89)	.07 ⁺ (.04)	-.02 (.07)	1.22** (.24)	1.75** (.42)	-1.99** (.54)	-4.06** (.14)
Pretreatment score	.46** (.14)	.39** (.12)	.43** (.09)	.40** (.14)	.69** (.10)	.72** (.09)	.24* (.12)	.34* (.12)
Boy	.29 (.29)	.77 (.72)	-.17 ⁺ (.10)	-.13 (.14)	-.75 (.97)	-.72 (.81)	-.37 (.83)	.76 (1.02)
Non-Hispanic black	.20 (.58)	-.01 (1.16)	.09 (.20)	.20 (.22)	1.11 (2.61)	-.26 (1.20)	.26 (.60)	1.02 (2.61)
Poverty-related risks	-.14 (.16)	-.20 (.31)	-.06 ⁺ (.04)	.01 (.07)	-.26 (.56)	-.01 (.41)	.16 (.35)	.31 (.60)
Single-parent family	.87 (1.01)	.34 (1.59)	-.08 (.19)	.00 (.15)	-.84 (1.26)	-.82 (.84)	.13 (.73)	-.41 (.84)
Speaking Spanish	-.60 (.45)	-.57 (.80)	.16 (.15)	-.22 (.24)	1.01 (1.84)	1.18 (1.25)	-.36 (1.00)	.39 (2.00)
Teacher personal stressors	-.59 (2.76)	-.65 (2.02)	-.12 (.11)	.05 (.14)	-.09 (2.01)	-1.22 (2.14)	.33 (1.04)	2.19 (3.84)
Teacher work stressors	.50 (2.80)	.03 (3.54)	.11 (.17)	.03 (.11)	1.04 (1.51)	.67 (1.82)	1.17 (2.06)	4.27 (3.84)
Teacher behavior management	-.15 (3.84)	1.32 (2.70)	-.04 (.15)	-.17 (.25)	-.28 (2.17)	-2.42 (3.45)	-3.66* (1.87)	-4.07* (1.87)
Teacher sensitivity	-1.70 (2.77)	-1.44 (3.28)	-.08 (.18)	.09 (.24)	.12 (1.37)	1.55 (2.71)	.65 (3.15)	1.12 (2.77)
Classroom overall quality	-1.84 ⁺ (1.02)	-.55 (5.62)	.15 (.16)	.11 (.16)	-.63 (3.87)	3.21* (1.55)	1.02 (1.84)	1.01 (2.77)
Class size	-.96 (1.41)	-.63 (1.88)	-.05 (.06)	.02 (.04)	.14 (.32)	-.17 (.38)	-.11 (.67)	-.09 (1.41)
No. adults in classroom	.71 (1.97)	2.19* (1.08)	.11 (.14)	-.03 (.14)	1.02 (1.89)	.31 (1.41)	1.71 (3.44)	4.38* (1.08)
Constant	4.91 (6.33)	5.01 (4.74)	1.63 (1.19)	.52 (1.01)	2.15 (9.52)	-.23 (4.61)	1.43 (2.70)	.59 (1.01)

Note.—BPI-I = Behavior Problem Index, internalizing scale; BPI-E = Behavior Problem Index, externalizing scale; A-I = attention and impulse control; Pos Emo = positive emotion; PPVT = Peabody Picture Vocabulary Test; Math = early math

skills. Standard errors are presented in parentheses. Results are combined from the estimates of five data sets generated by multiple imputation.

⁺
 $p < .10.$

^{*}
 $p < .05.$

^{**}
 $p < .01.$

Fig. 1.
Average teacher-training hours by class

Table 1

Descriptive Statistics by Treatment Condition and Matching Status

	Full Sample	Full Control (a)	Teacher Training				MHC Class Visits				Individual MHC	
			Low Dosage		High Dosage		Low Dosage		High Dosage		Treated (b)	Matched Control (c)
			Treated (b)	Matched Control (c)	Treated (b)	Matched Control (c)	Treated (b)	Matched Control (c)				
Child and family characteristics:												
Sample size	602	294	175	175	114	114	123	123	185	185	137	137
Boy	.47	.44	.54 ⁺	.48	.50	.45	.52 ⁺	.47	.50	.48	.62 ^{**}	.55
Non-Hispanic black	.66	.65	.81 ^{**}	.84	.39 ^{**}	.51 ⁺	.79 ^{**}	.85	.58	.65	.71	.70
Poverty-related risks	1.08	.99	1.26 ^{**}	1.17	.91	1.02	1.44 ^{**}	1.30	.98	.93	1.21 [*]	1.10
Single-parent family	.69	.67	.75 [*]	.77	.61	.64	.74	.75	.68	.73	.69	.72
Speaking Spanish	.23	.25	.17 [*]	.15	.28	.34	.16 ⁺	.17	.24	.26	.18 ⁺	.20
Pretreatment scores:												
BPI internalizing scale	2.24	1.99	2.16	2.03	3.06 ^{**}	3.08	2.01	1.98	2.78 ^{**}	2.21	2.89 ^{**}	2.19 ⁺
BPI externalizing scale	5.72	5.17	5.91	5.82	7.08 ^{**}	6.60	6.32 ⁺	5.77	6.20 [*]	5.99	8.19 ^{**}	7.74
Attention and impulse control	2.15	2.10	2.17	2.19	2.21	2.17	2.15	2.10	2.24 [*]	2.18	2.15	2.14
Positive emotion	1.91	1.88	1.95	2.01	1.93	1.82	1.97	2.01	1.93	1.89	1.98	1.98
PPVT	10.33	9.98	10.22	10.10	11.35 ^{**}	10.77	10.81 ⁺	10.35	10.56	10.52	10.63	9.90
Early math skills	7.07	6.70	7.49 [*]	7.27	7.43	7.62	7.64 [*]	7.17	7.29	7.03	6.99	7.02
Teacher and classroom characteristics:												
Sample size	35	17	10	10	7	7	8	8	10	10	18	17
Teacher personal stressors	2.59	2.45	3.08 ⁺	2.75	2.33	2.02	3.04 ⁺	2.94	2.48	2.42	2.73	2.45
Teacher work stressors	1.23	1.30	1.10	1.27	1.21	1.15	1.48	1.46	.92	1.25	1.17	1.30
Teacher behavior management	4.92	5.18	4.21 [*]	4.52	5.30	5.40	4.21 [*]	4.52	5.04	5.10	4.67	5.18
Teacher sensitivity	4.82	5.11	4.26 ⁺	4.51	4.91	5.21	4.54	4.62	4.57	5.06	4.56	5.11
Classroom overall quality	4.75	4.97	4.08 ^{**}	4.35	5.22	5.09	4.81	4.95	4.34 [*]	4.67	4.55	4.97
Class size	16.03	16.00	16.80	15.70	14.71	14.43	15.13	15.25	16.80	16.00	16.06	16.00

	Teacher Training				MHC Class Visits							
	Low Dosage		High Dosage		Low Dosage		High Dosage					
	Full Sample	Full Control (a)	Treated (b)	Matched Control (c)	Treated (b)	Matched Control (c)	Treated (b)	Matched Control (c)	Treated (b)	Matched Control (c)		
No. adults in classroom	2.29	2.21	2.03	2.28	2.86 ⁺	2.46	1.95	2.06	2.70 ⁺	2.53	2.37	2.21

Note.—MHC =mental health consultant; BP1 =Behavior Problem Index (internalizing and externalizing scales); PPVT =Peabody Picture Vocabulary Test III. Means in respective samples: child and family covariates in child-level samples and teacher and classroom covariates in class-level samples; *p*-values represent results for two-tailed *t*-statistics testing the Full Control column's (col. *a* i.e., control group before matching) differences from the Treated columns' (cols. *b*), as well as the Treated columns' (cols. *b*), as well as the Treated columns' (cols. *b*) mean differences from the Matched Control columns (cols. *c*; i.e., control groups after matching). Significance levels for the Full Control column's (*a*) differences from the Treated columns (*b*) are presented in the Treated columns; levels for differences between the Treated (*b*) and Matched Control (*c*) columns are presented in the Matched Control columns.

⁺ *p* < .10.
^{*} *p* < .05.
^{**} *p* < .01.

Table 2

Program Effects of CSRP Intervention on Children's School Readiness

	ITT	Teacher Training		MHC Class Visits		MHC for Children
		Low Dosage	High Dosage	Low Dosage	High Dosage	
Teacher-reported behavior problems:						
BPI internalizing scale	−1.81 ^{**} (.43)	−1.36 ^{**} (.46)	−1.99 ^{**} (.54)	−1.12 ^{**} (.39)	−1.92 ^{**} (.52)	−1.61 ^{**} (.30)
BPI externalizing scale	−2.92 [*] (.92)	−2.27 [*] (.89)	−4.06 ^{**} (.87)	−2.10 ^{**} (.33)	−3.33 ^{**} (.79)	−2.41 ^{**} (.54)
Observer-reported social-emotional skills:						
Attention and impulse control	.20 [*] (.08)	.07 (.06)	.49 ^{**} (.10)	.14 ^{**} (.03)	.43 ^{**} (.12)	.31 ^{**} (.08)
Positive emotion	−.01 (.08)	−.02 (.07)	.40 ^{**} (.09)	−.01 (.12)	.31 ^{**} (.08)	.18 [*] (.05)
Observer-rated cognitive development:						
PPVT	1.46 [*] (.61)	1.22 ^{**} (.45)	3.70 ^{**} (.89)	1.10 ⁺ (.63)	2.59 ^{**} (.55)	2.37 ^{**} (.70)
Early math skills	2.21 ^{**} (.52)	1.75 ^{**} (.42)	3.74 ^{**} (.74)	1.37 ^{**} (.30)	2.78 ^{**} (.80)	2.15 [*] (.96)

Note.—CSRP = Chicago School Readiness Project; ITT = intention to treat; MHC = mental health consultant; BPI = Behavior Problem Index; PPVT = Pea-body Picture Vocabulary Test III. Standard errors are presented in parentheses. Results are combined from the estimates of five data sets generated by multiple imputation.

⁺ $p < .10$.

^{*} $p < .05$.

^{**} $p < .01$.