Symposium Title: Testing questions of mechanism within early educational interventions that used experimental or quasiexperimental design: Four alternative approaches

Having recognized and embraced the importance of experimental (e.g. cluster-RCT) and quasi-experimental (e.g. regression discontinuity) designs for being able to make causal inferences, researchers examining educational effectiveness now face a second set of equally serious intellectual challenges. Estimates of impact from those experiments and quasiexperiments provide “causal description,” allowing us to draw the conclusion that children’s educational outcomes (in the short-term, at least) are significantly improved with increased investment in early childhood education, or simply that a given early educational intervention works. But such designs do not help us with an equally important goal of “causal explanation,” on elucidating how intervention works (see Shadish, Cook & Campbell, 2005). As one of our submitted papers so succinctly puts it: “Unpacking the black box of intervention effects by testing process mediators is a critical step for identifying the key features of programs that are likely to promote successful replication and sustainability as well as for supporting program improvement efforts. To date however too little attention has been paid to the methodological challenges associated with drawing causal inferences from experimental evaluations of programs that hold multi-level theories of change.”

In the following four papers, we contrast several different approaches taken by seasoned research teams with extensive experience in experimental designs to address this second, increasingly pressing intellectual question. The four papers offer complementary perspectives on ways to identify and test models of the mechanisms that may be responsible for observed change in children’s academic and socioemotional outcomes in the context of early educational intervention.

The first papers focuses closely on young children’s academic, behavioral, and neurocognitive gain within the intervention year: Presented by Weiland and Yoshikawa, this paper outlines new findings using RD design that detect the impact of children’s enrollment in preK on their school readiness. Importantly, this paper highlights the role of a key child-level mediator, i.e., their executive function.

The symposium’s second paper will offer important insights regarding mechanisms at work across multiple RCTs by using meta-analysis to consider key features of early educational interventions that may serve as potential levers of change in child outcomes. Duncan and colleagues will present recent results of their meta-analysis of early educational interventions where they consider experimental studies from 1960 to 2007 (rather than students or schools) as the units of analysis, capitalizing on variation in intervention design to estimate the role of optimal starting age, optimal program duration and persistence of impacts. This paper will serve as an excellent “bridge” to the 3rd and 4th papers, where modeling longer-term impacts of educational intervention will be considered.
Specifically, the third paper (presented by Lowenstein and colleagues) will consider ways that kindergarten represents an understudied second, nonrandomly-assigned treatment condition that must be modeled carefully in order to understand the ways that preschool-treatment impacts may be sustained or attenuated over time. This paper will deploy lagged autoregressive models using HLM and SEM to estimate to role of positive versus chaotic school climate as a second “treatment” predicting kindergarten outcomes. Resulting estimates from those analyses will be compared to other long-term impact estimates obtained from propensity score matching analyses (previously completed by other members of the research team, with the same data set).

The fourth paper will aid us in modeling longer-term effects of early intervention by using a cluster-randomized design in 18 schools to consider the roles of both school- and child-level change. Jones and colleagues will compare results from standard tests of mediation (Baron & Kenny, 1986) to an instrumental variables approach, conducting analyses across a wide range of domains of classroom process and child outcome including children’s social cognitive processes, social-emotional symptomatology, academic achievement, and aggressive versus socially competent functioning. This rich data source (including 6 waves of data collection across 3 years) will allow for sophisticated models of students’ growth trajectories of both academic and socioemotional skill as well as for changes in multiple elements of school climate across time.
Title: The Impact of an Urban Universal Public Prekindergarten Program on Children’s Early Numeracy, Language, Literacy, and Executive Function Outcomes

Author(s): Christina Weiland and Hirokazu Yoshikawa
Background / Context:  
Description of prior research and its intellectual context.

According to the literature, high-quality early childhood education equips children with the cognitive and academic skills required to be successful in elementary school and beyond (Gilliam, 2005; National Forum on Early Childhood Program Evaluation, 2007). Landmark studies of particularly intensive interventions have found positive impacts that last into adulthood and that are highly cost effective (Campbell, Ramsey, Pungello, Sparling & Miller-Johnson, 2002; Currie, 2001; Karoly, Kilborn & Cannon, 2005; Schweinhart, Barnett & Belfield, 2005; Yoshikawa, 1995), with estimated cost-benefit ratios ranging from about $3 to $12 for every $1 spent (Heckman, Moon, Pinto, Savelyev & Yavitz, 2010; Temple & Reynolds, 2007).

It is unclear in the literature, however, whether or how such impacts can be maintained at scale. Generalizing the results of small-scale experiments to at-scale programs raises a host of problems, as small-scale programs are generally more resource-intensive and have extensive involvement of the creators of the intervention that simply is not possible at scale (Murnane & Willett, 2010; Yoshikawa, Rosman & Hsueh, 2002). This may explain why many large-scale preschool programs often have weaker effects than model programs (Barnett, 1995; Currie & Thomas, 1995).

Research on an increasingly common large-scale preschool model – state-funded prekindergarten – has lagged behind its policy scale up. The number of states offering such programs increased from 10 in 1980 to 38 in 2009 (Gormley, Gayer, Phillips, & Dawson, 2005; NIEER, 2009). Yet, only a handful of studies have examined the causal impacts of these programs (Gormley, Gayer, Phillips, & Dawson, 2005; Hustedt, Barnett, Jung & Goetze, 2009; Hustedt, Barnett, Jung & Thomas, 2007; Wong, Cook, Barnett, & Jung, 2007). Encouragingly, these studies have found small to moderate positive impacts on children’s language, literacy and numeracy skills. However, no study to date has examined the causal impact of a large-scale state-funded public prekindergarten program on child executive functioning, a domain many developmentalists consider to be an important component of school readiness.

Existing studies of large-scale state-funded public prekindergarten programs also do not fully address the question of under what conditions these programs can achieve impacts at scale, as there was no consistent curriculum in place in any of the programs examined in this literature (Gormley, Gayer, Phillips, & Dawson, 2005; Hustedt, Barnett, Jung & Goetze, 2009; Hustedt, Barnett, Jung & Thomas, 2007; Wong, Cook, Barnett, & Jung, 2007). Theory and some empirical research suggest that implementing an intentional curriculum may improve child outcomes by helping to ensure program quality, by keeping children engaged and challenged and by building specific skills targeted by the curriculum (Klein & Knitzer, 2006; NAEYC & NAECS/SDE, 2003). Finally, existing prekindergarten studies have not explored sensitivity of results to some common methodological challenges, including possible non-comparability of treatment and control children due to differential time to join and attrit from the intervention.

Purpose / Objective / Research Question / Focus of Study:  
Description of the focus of the research.

We add to and extend the emerging evidence base of the effects of public preschool
programs on child school readiness. Using a quasi-experimental, Regression Discontinuity (RD) design, we estimate the impacts of a year of a universal preschool program on children’s early numeracy, language, literacy, and executive function skills, both for the overall population and for several subgroups. While we find impacts similar to those reported in other public prekindergarten RD studies, we make a unique contribution to the literature, as ours is the first causal study of a universal prekindergarten program in which a uniform curricula was in place across the district and in which we have information about the type of care experienced by control children during the treatment year. At SREE, we will also present results from robustness checks new to this literature, including sensitivity of results to differential time to join and attrit from the intervention.

Setting:
Description of the research location.

Research took place in a large urban public school district in the Northeast. All prekindergarten programs were located in public elementary schools.

Population / Participants / Subjects:
Description of the participants in the study: who, how many, key features or characteristics.

In Fall 2009, children in a citywide 4-year-old prekindergarten program and all children who attended the program in the previous year were eligible for the our study, so long as they were placed in non-special education only classrooms (criterion 1), they did not permanently withdraw from the district before October 1 of their prekindergarten year (criterion 2), and they were enrolled in school at least seven days (criterion 3). Special education-only classrooms were excluded due to concerns about the appropriateness of the assessment battery for children who were not mainstreamed. Criteria (2) and (3) were employed to filter out children who register for prekindergarten but never show up or show up for so few days that they were exposed to very little of the program and could not be located to be assessed.

For a child to participate in the study, the principal, classroom teacher, and parent/guardian of the child had to consent to participate. Out of 79 elementary schools with eligible children, 12 principals declined to participate (15%). Over 93% of eligible teachers in participating schools agreed to participate. Within participating classrooms in the 67 participating schools, 69.2% of eligible children returned consent forms, for a total sample size of 2,018. This represents 55% of eligible children in the district. As evident in Table 1, the final sample of participating children is racially and linguistically diverse.

Intervention / Program / Practice:
Description of the intervention, program or practice, including details of administration and duration.

The program is universal in design, as any child within the city who turns four by September 1 in a given year can apply for the program. However, there is more demand than there is supply, with a lottery system determining which students are allotted prekindergarten seats. Currently, approximately 35% of the city’s 4-year-old population attends the prekindergarten program. All prekindergarten classrooms in the districts are staffed with at least one teacher with at least a B.A. and one paraprofessional (adult-child ratio is about 1:10). Prekindergarten teachers are paid on the same scale as K-12 teachers. Intending to promote
classroom quality, the district implemented the literacy curriculum Opening the World of Learning (OWL) (Schickedanz & Dickinson, 2005) and the mathematics curriculum Building Blocks (Clements & Sarama, 2007a) system-wide in 4-year-old classrooms in the 2007-2008 school year. Treatment children in our study attended the program in the 2008-2009 school year, while control children attended the program in the 2009-2010 school year.

Research Design:
Description of research design (e.g., qualitative case study, quasi-experimental design, secondary analysis, analytic essay, randomized field trial).

We employ a Regression Discontinuity Design to obtain causal estimates, with the birthday cutoff for entry into the program in a given year as the source of exogeneity. Importantly, the district strictly enforces the cutoff; in recent years, no child has been admitted into the program when their birthdate suggests they should not. Approximately 94% of children offered a spot in the 4-year-old program in 2008-2009 enrolled in the program for at least one day.

Data Collection and Analysis:
Description of the methods for collecting and analyzing data.

Children were tested by study-trained child assessors. All assessors were college educated and approximately one third held masters degrees. On average, the battery of tests took approximately 45-50 minutes to administer. Testers were instructed to test children in one session if possible but to divide the session into smaller segments if children showed signs of fatigue. Because of the session length, we randomly varied the order of the tests to limit the possibility of systematically biasing results due to child fatigue. The assessors visited classrooms in Fall 2009, as close to the start of the school year as possible. Measures of pre-literate, numeracy and executive function skills were collected. See Table 2 in Appendix B for a list of measures used in our study.

Our implementation of the RD framework is guided by the advice of Lee and Lemieux (2009), by the strategy and organization of Wong et al. (2007), and by the recently released What Works Clearinghouse guidelines (Schochet et al., 2010). We first conduct a graphical analysis of the relationship between the outcome and smoothed function of child age on either side of the cutoff. As shown in Figure 1, on either side of the cutoff, we superimpose a linear regression line and a smoothed, locally weighted non-parametric regression line on a scatterplot of the raw data. These graphs give some indication of functional form, as well as whether there is indeed a “jump”, or difference between the two groups, at the cutoff. Second, because identifying the correct functional form of the continuous assignment variable is one of the chief challenges in RD analysis (Lee & Lemieux, 2009), we fit a series of regression model specifications, including polynomials, interaction terms and non-parametric models, to the raw data. We compare relative fit across models and purposely overspecify the model as a robustness check. Although less efficient than underspecifying, overspecifying yields more unbiased coefficients (Trochim, 1984) and has been used as a strategy in other early childhood RD designs (Gormley et al., 2005; Wong et al., 2007).

Our primary equation for fitting regression models is as follows:

\[ \text{OUTCOME}_g = \beta_0 + \beta_1 \text{TREAT}_g + \beta_2 \text{AGE}_g + \beta_3 \text{TREAT}_g \times \text{AGE}_g + \delta \text{AGE}_g + (\alpha_g + \delta_g) \]  

(1)

where \text{OUTCOME} is a child-level test score, \text{TREAT} is a dummy variable that takes on the value of 1 if the student’s birthday is on or before September 1, 2004 and the value of 0 if not, \text{AGE} is
a smooth function of the student’s age measured in days and centered on the September 1 birthday cutoff, $TREAT*AGE$ is an interaction term that allows the effect to vary on either side of the cutoff, $^*X$ is a vector of student demographic covariates, $^\dagger\epsilon$ is the error term associated with students and $^\delta\delta$ is the error term associated with classrooms. Subscript $i$ denotes students and subscript $j$ denotes classrooms. In all regression models, we adjust standard errors for clustering at the classroom level. In all regression models, we use multiple imputation (with 50 imputations) to account for missing data in accordance with Graham (2009).

We extend our approach to examine effects by subgroups by fitting separate regressions for each subgroup. Subgroups of interest are defined by home language (English, Spanish, other), race (Black, Latino, White, Asian), free/reduced lunch, gender, special needs status, and pre-program childcare experience (Head Start, private, family daycare, none, public).

We also perform a series of robustness checks. Threats to the internal validity of our results include: 1) treatment misallocation at the cutoff; 2) non-smooth or discontinuous variation of observed and unobserved student characteristics around the cutoff; 3) discontinuities in the outcomes at points other than the cutoff; 4) incorrect specification of the functional form; and 5) sensitivity of results to choice of bandwidth. All of these threats could result in either an over- or under-estimation of the true impact of the treatment. We do not have space in this proposal to discuss how we address each threat specifically; however, our robustness checks follow current best practices in the RD literature (Imbens & Lemieux, 2008; Lee & Lemieux, 2009).

At SREE, we will present results from several robustness checks new to the prekindergarten RD literature. For example, we will explore sensitivity of results to differential time to enter the treatment and control groups; in our study, treatment children had a full school year to enter prekindergarten (2008-2009), while control children had only a few months (Fall 2009). To address this challenge, we will examine the sensitivity of results to different definitions of the sample (i.e. children who were enrolled the full time vs. those enrolled at least one month vs. those enrolled at least one day). To equalize attrition time, we will obtain administrative records for the control group that cover the same amount of time as the treatment group (i.e. through December of the kindergarten year) and refit models without study control children who later attrited from the district. Finally, we will address the methodological challenge posed by different start rules in commonly used early childhood assessments for children of different ages.

Findings / Results:
Description of the main findings with specific details.

As evident from Table 3, we found significant (p<0.05), small to moderate and positive effect sizes on all assessments. Effect sizes for numeracy skills as measured by the Applied Problems assessment were largest (0.61), with effect sizes for pre-reading and reading skills (0.55) and vocabulary (0.44) somewhat smaller. Executive functioning effect sizes were in the small range but all positive: 0.25 (inhibitory control), 0.27 (cognitive flexibility), and 0.26 (working memory). $^\ddagger$ We also found significant positive impacts on most outcomes for most subgroups, which is notable given our reduced power to detect effects within subgroups (exceptions were White children and children with special needs). Results were robust across

$^*$ Allowing the slope to vary on either side of the cutoff is necessary to estimate unbiased estimates of impact at the cutoff (Imbens & Lemieux, 2008).

$^\dagger$ Including demographics as covariates increases the precision of estimates (Imbens & Lemieux).

$^\ddagger$ We do not present effect sizes for all outcomes or all subgroups, as some analyses are currently underway. At SREE in March, we will present impacts for all assessments shown in Table 2 and for all subgroups of interest.

2011 SREE Conference Abstract Template
multiple bandwidths and model specifications, and in other standard RD robustness checks, we find no reason to doubt our findings.

**Conclusions:**
**Description of conclusions, recommendations, and limitations based on findings.**

Our results add to the growing literature on the causal effects of large-scale state-funded prekindergarten programs. We find that a universal publicly funded prekindergarten program has positive impacts on child early numeracy, language, literacy, and executive function skills. Results for executive function are particularly of interest, given that ours is the first study to provide causal evidence of the impact of a large-scale publicly funded prekindergarten program on this important developmental domain.

In Table 4, we place our main impact language, literacy, and early numeracy results in the context of other RD prekindergarten studies. Our effect sizes are larger than those achieved in any RD prekindergarten study to date. We believe there are several possible reasons for this. First, stronger impacts in our study might be explained by differences in the respective counterfactuals. We can only speculate regarding this possibility, as ours is the only RD study to date with information on the counterfactual experienced by the control group and as we have information regarding only the type, not the quality, of control group care settings.

Second, our sample includes only children whose parents gave active consent for them to participate, while many of the other examined contexts required passive consent. This difference raises external validity issues; while our estimates are internally valid, we can generalize our estimates only to children whose parents consented to participate. Thus, it may be that the our sample and samples in other prekindergarten RD studies are not exactly comparable and that this difference fully or partially explains the larger effects estimated in our context.

Finally, as mentioned previously in this paper, ours is the only RD prekindergarten context in which there was a uniform curriculum in place. Stronger effects thus could at least partially be a function of the chosen math and literacy curricula and the district’s implementation/ professional development supports. This explanation is consistent with theory and some empirical research that suggests that implementing an intentional curriculum may improve child outcomes by helping to ensure program quality, by keeping children engaged and challenged and by building specific early skills, in domains such as literacy, numeracy, socio-emotional, or self-regulation skills (Klein & Knitzer, 2006; NAEYC & NAECS/SDE, 2003). The overall evidence base for the efficacy of preschool curricula is mixed, as many randomized trials have found small or null effects of curricula on children’s math and literacy skills (Judkins et al, 2008; Preschool Curriculum Evaluation Research Consortium, 2008; WWC, n.d.). However, given promising results to date for the specific curricula used in the district (Clements & Sarama, 2007b; Clements & Sarama, 2008; Clements, Sarama, Spitler, Lange & Wolfe, in press; Pearson, n.d.), stronger effects could be explained at least partially by the use of a uniform curricula.

By March 2011, we will also explore a number of methodological issues that have not yet been addressed in prekindergarten RD studies. Chief among these, we will examine the sensitivity of our results to the definition of the sample, due to differential enrollment and attrition timeframes for the treatment and control groups. We are currently awaiting data from the district for these checks but they will be complete by March 2011. Finally, we also explore the sensitivity of our results to different start rules on the PPVT-III for treatment and control children. Consistent with test guidelines, in our study, 5-year-old children started the PPVT-III one set ahead of 4-year-old children.
Appendices
Not included in page count.

Appendix A. References


Table 1: Descriptive characteristics of sample

<table>
<thead>
<tr>
<th>Variable</th>
<th>Overall (N=2018)</th>
<th>Born before cutoff (N=969)</th>
<th>Born after cutoff (N=1049)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attendance zone is the North Zone</td>
<td>0.28</td>
<td>0.29</td>
<td>0.26</td>
</tr>
<tr>
<td>Attendance zone is the East Zone</td>
<td>0.44</td>
<td>0.45</td>
<td>0.44</td>
</tr>
<tr>
<td>Attendance zone is the West Zone</td>
<td>0.28</td>
<td>0.26</td>
<td>0.30</td>
</tr>
<tr>
<td>English only home language</td>
<td>0.50</td>
<td>0.48</td>
<td>0.53</td>
</tr>
<tr>
<td>Spanish home language</td>
<td>0.27</td>
<td>0.28</td>
<td>0.27</td>
</tr>
<tr>
<td>Other home language</td>
<td>0.22</td>
<td>0.24</td>
<td>0.20</td>
</tr>
<tr>
<td>Black</td>
<td>0.27</td>
<td>0.28</td>
<td>0.25</td>
</tr>
<tr>
<td>White</td>
<td>0.18</td>
<td>0.18</td>
<td>0.19</td>
</tr>
<tr>
<td>Hispanic</td>
<td>0.41</td>
<td>0.39</td>
<td>0.42</td>
</tr>
<tr>
<td>Asian</td>
<td>0.11</td>
<td>0.11</td>
<td>0.11</td>
</tr>
<tr>
<td>Other race/ethnicity</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>Special Needs</td>
<td>0.09</td>
<td>0.11</td>
<td>0.08</td>
</tr>
<tr>
<td>Free/reduced lunch receipt</td>
<td>0.69</td>
<td>0.72</td>
<td>0.66</td>
</tr>
<tr>
<td>Male</td>
<td>0.51</td>
<td>0.52</td>
<td>0.50</td>
</tr>
<tr>
<td>Previously attended family daycare</td>
<td>0.07</td>
<td>0.08</td>
<td>0.06</td>
</tr>
<tr>
<td>Previously attended Head Start</td>
<td>0.16</td>
<td>0.16</td>
<td>0.16</td>
</tr>
<tr>
<td>Did not attend any care program previously</td>
<td>0.34</td>
<td>0.34</td>
<td>0.33</td>
</tr>
<tr>
<td>Previously attended public preschool</td>
<td>0.11</td>
<td>0.11</td>
<td>0.11</td>
</tr>
<tr>
<td>Previously attended private center care</td>
<td>0.33</td>
<td>0.31</td>
<td>0.35</td>
</tr>
</tbody>
</table>

*Note: one child born after the cutoff is missing all information in this table. 76 children (4% of sample) are missing pre-program care data.
Table 2: Child Assessment Battery

<table>
<thead>
<tr>
<th>Name of Assessment</th>
<th>Domain</th>
<th>Specific construct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peabody Picture Vocabulary Test – III (PPVT-III) (Dunn &amp; Dunn, 1997)</td>
<td>Language</td>
<td>Receptive vocabulary</td>
</tr>
<tr>
<td>Woodcock-Johnson Letter-Word Identification (Woodcock, McGrew &amp; Mather, 2001)</td>
<td>Pre-Literacy</td>
<td>Pre-reading and reading</td>
</tr>
<tr>
<td>Research-based Elementary Math Assessment (REMA) (Clements, Sarama &amp; Liu, 2008)</td>
<td>Numeracy</td>
<td>Comparing/ordering, verbal counting/counting strategies, arithmetic, number recognition and subitizing, geometric, measuring and patterning capacities</td>
</tr>
<tr>
<td>Forward Digit Span (Gathercole &amp; Pickering, 2000; Wechsler, 1986)</td>
<td>Executive function</td>
<td>Working memory (phonological loop)</td>
</tr>
<tr>
<td>Backward Digit Span (Gathercole &amp; Pickering, 2000; Wechsler, 1986).</td>
<td>Executive function</td>
<td>Working memory (central executive)</td>
</tr>
<tr>
<td>Dimensional Change Card Sort (DCCS) (Frye, Zelazo &amp; Palfai, 1995)</td>
<td>Executive function</td>
<td>Attention Shifting</td>
</tr>
<tr>
<td>Pencil Tapping (Diamond &amp; Taylor, 1996)</td>
<td>Executive function</td>
<td>Inhibitory control</td>
</tr>
</tbody>
</table>
Table 3: Estimated effect sizes for sample children who participated in the preschool program across child outcomes and by subgroup

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Full Sample</td>
<td>0.440***</td>
<td>0.554***</td>
<td>0.609***</td>
<td>0.249**</td>
<td>0.257*</td>
<td>0.269**</td>
</tr>
<tr>
<td>Student-level demographics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>0.397**</td>
<td>0.637**</td>
<td>0.535***</td>
<td>0.395*</td>
<td>0.009</td>
<td>0.042</td>
</tr>
<tr>
<td>White</td>
<td>0.098</td>
<td>-0.047</td>
<td>0.035</td>
<td>-0.049</td>
<td>0.001</td>
<td>0.011</td>
</tr>
<tr>
<td>Hispanic</td>
<td>0.478***</td>
<td>0.217***</td>
<td>0.164***</td>
<td>0.117**</td>
<td>0.018~</td>
<td>0.067**</td>
</tr>
<tr>
<td>Asian</td>
<td>0.644~</td>
<td>0.127</td>
<td>0.266*</td>
<td>0.010</td>
<td>0.028~</td>
<td>0.141*</td>
</tr>
<tr>
<td>Free/reduced lunch</td>
<td>0.487***</td>
<td>0.173***</td>
<td>0.165***</td>
<td>0.113***</td>
<td>0.014~</td>
<td>0.072**</td>
</tr>
<tr>
<td>Full price lunch</td>
<td>0.311*</td>
<td>0.081</td>
<td>0.083*</td>
<td>-0.005</td>
<td>0.018*</td>
<td>0.030</td>
</tr>
<tr>
<td>Male</td>
<td>0.439**</td>
<td>0.166***</td>
<td>0.144***</td>
<td>0.083*</td>
<td>0.026**</td>
<td>0.046</td>
</tr>
<tr>
<td>Female</td>
<td>0.445**</td>
<td>0.130**</td>
<td>0.142***</td>
<td>0.076~</td>
<td>0.006</td>
<td>0.076**</td>
</tr>
<tr>
<td>Special needs</td>
<td>0.264</td>
<td>0.006</td>
<td>0.079</td>
<td>-0.013</td>
<td>0.009</td>
<td>0.094</td>
</tr>
<tr>
<td>No special needs</td>
<td>0.468***</td>
<td>0.174***</td>
<td>0.148***</td>
<td>0.100***</td>
<td>0.017*</td>
<td>0.055*</td>
</tr>
<tr>
<td>Home language</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>0.201~</td>
<td>0.108*</td>
<td>0.079**</td>
<td>0.021</td>
<td>0.013~</td>
<td>0.006</td>
</tr>
<tr>
<td>Spanish</td>
<td>0.687***</td>
<td>0.190***</td>
<td>0.204***</td>
<td>0.146**</td>
<td>0.021~</td>
<td>0.102**</td>
</tr>
<tr>
<td>Other</td>
<td>0.724**</td>
<td>0.200***</td>
<td>0.216***</td>
<td>0.137*</td>
<td>0.019</td>
<td>0.131**</td>
</tr>
</tbody>
</table>

Note: Effect sizes are from models with linear specification of the distance from the age cutoff and with robust SEs with correction for clustering at the classroom-level. Subgroup models also include an interaction between the dichotomous preschool variable and distance from age cutoff. Full sample models include student-level covariates. Effect sizes calculated using the SD of the control group. ***p<0.001; **p<0.01; *p<0.05; ~p<0.10
Table 4: Comparison of effect sizes across RD prekindergarten studies

<table>
<thead>
<tr>
<th></th>
<th>PPVT-III</th>
<th>Letter Word Identification</th>
<th>Applied Problems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Present study</td>
<td>0.44***</td>
<td>0.63***</td>
<td>0.61***</td>
</tr>
<tr>
<td>Tulsa</td>
<td>--</td>
<td>0.80***</td>
<td>0.38*</td>
</tr>
<tr>
<td>Michigan</td>
<td>-0.16</td>
<td>--</td>
<td>0.47*</td>
</tr>
<tr>
<td>New Jersey</td>
<td>0.36*</td>
<td>--</td>
<td>0.23*</td>
</tr>
<tr>
<td>South Carolina</td>
<td>0.05</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>West Virginia</td>
<td>0.14</td>
<td>--</td>
<td>0.11</td>
</tr>
<tr>
<td>Oklahoma</td>
<td>0.29*</td>
<td>--</td>
<td>0.35</td>
</tr>
<tr>
<td>New Mexico, Year 1</td>
<td>0.35+</td>
<td>--</td>
<td>0.38+</td>
</tr>
<tr>
<td>New Mexico, Year 2</td>
<td>0.25+</td>
<td>--</td>
<td>0.50+</td>
</tr>
<tr>
<td>New Mexico, Year 3</td>
<td>0.17+</td>
<td>--</td>
<td>0.43+</td>
</tr>
</tbody>
</table>

***p<0.001; **p<0.01; *p<0.05 ; + results statistically significant but level of significance not reported.
Note: Effect sizes calculated using the SD of the control group.
Citations: Tulsa (Gormley, Gayer, Phillips, & Dawson, 2005); MI, NJ, SC, WV, OK (Wong et al., 2007); NM (Hustedt, Barnett, Jung & Goetze, 2009).
Figure 1: Scatterplot of Peabody Picture Vocabulary Test-III scores, with super-imposed fitted linear regression and lowess regression lines.
Title: Timing issues with early childhood education programs: how effect sizes vary by starting age, program duration and persistence of effects

Author(s):

Greg J. Duncan, University of California, Irvine Department of Education (gduncan@uci.edu)

James A. Leak, University of California, Irvine Department of Education (jleak@uci.edu)

Weilin Li, University of California, Irvine Department of Education (weilinl@uci.edu)

Katherine Magnuson, University of Wisconsin-Madison (kmagnuson@wisc.edu),

Holly Schindler, Center for the Developing Child, Harvard University (holly_schindler@harvard.edu)

Hiro Yoshikawa, Harvard Graduate School of Education (yoshiki@gse.harvard.edu).
Three interesting timing issues for early childhood education program are intertwined: optimal starting age, optimal program duration and the persistence of impacts. In the case of starting age, “skill begets skill” human capital production models (Cunha & Heckman, 2007) provide a justification for very early intervention, since boosting skills can improve the productivity of later investments. For our country’s universal K-12 schooling “intervention,” this logic suggests that children most ready for kindergarten are best able to profit from the next 13+ years spent in school. But the same logic may apply to preschool investments. If children most ready for an age 4 pre-K program profit the most from them, it may be better to begin boosting children’s skills at age 3 or even earlier to increase the productivity of the age-4 programs.

On the other hand, early investments not followed up with high-quality subsequent investments may produce only ephemeral impacts. In the case of the Perry Preschool intervention (Schweinhart et al., 1993), the large cognitive impacts estimated shortly after the completion of the program had completely disappeared by age 8, although impacts on achievement, attainment and, eventually, crime and earnings persisted.

With starting age and follow-up length held constant, one would expect that longer programs would produce bigger impacts. And indeed, the Gorey (2001) meta-analysis reported that programs with durations in excess of 3 years had larger effects than 1 or 2 year-duration programs. Other studies have largely come to similar conclusions (Barnett & Lamy, 2006).

The focus of this paper centers around timing associated with early childhood education programs and interventions using meta-analytic methods. At any given assessment age, a child’s current age equals starting age, plus duration of program, plus years since program ended. Variability in assessment ages across our studies should enable us to identify the separate effects of all three time-related components. Combining these three components within the same analysis allows us to propose the following research questions:

1) When is the optimal timing for an intervention during the prenatal to age 5 period?
2) Should early education programs begin shortly after birth or is program initiation at age 3 or 4 just as beneficial for children’s learning?
3) Do early programs, which are introduced when children are developing on very different schedules, fade out more quickly than programs introduced later in early childhood?
4) Do longer-duration programs have less fade-out than shorter programs?

The project is a meta-analysis of evaluation studies of early childhood education programs conducted in the United States and its territories between 1960 and 2007.
Population / Participants / Subjects:
Description of the participants in the study: who, how many, key features or characteristics.

The population of interest is children enrolled in early childhood education programs between the ages of 0 and 5 and their control-group counterparts. Since the data come from a meta-analysis, the population for this study is drawn from many different studies with diverse samples.

Intervention / Program / Practice:
Description of the intervention, program or practice, including details of administration and duration.

Again, as a result of the nature of meta-analysis there is not a particular intervention or program being studied. Instead, we analyze the effects of multiple early childhood education programs including Head Start, Perry Preschool, and many other interventions and early childhood education programs for children ages 0-5.

Research Design:
Description of research design (e.g., qualitative case study, quasi-experimental design, secondary analysis, analytic essay, randomized field trial).

The research design of this study is meta-analysis. Instead of students or schools, we use prior studies as our unit of analysis. Meta-analysis allows researchers to gather information about prior studies and then estimate effect sizes of various components of the combined research studies. Effect sizes are expressed in standard deviation units and allow for the aggregation of effects of many programs into an overall program effect (Cooper and Hedges, 2009). Average effect sizes are compared across studies for differences in study design components, domain variables, and other study components.

Data Collection and Analysis:
Description of the methods for collecting and analyzing data.

This project involves a multi-step data collection and evaluation process for determining what studies will be included in the meta-analytic database. The first step was to conduct a comprehensive search of the literature from 1960-2007. The meta-analysis project started in 2007, thus the cut off date for inclusion in the database was 2007. The National Forum on Early Childhood Policy and Programs, which is the umbrella organization for our own work, was able to take advantage of a meta-analytic database compiled by Abt Associates, Inc. and the National Institute for Early Education Research (NIEER), which included early childhood intervention studies from 1960-2003 (Camilli et al., 2010; Jacob, Creps & Boulay, 2004; Layzer, Goodson, Bernstein & Price, 2001). This previous meta-analysis yielded 624 previously coded studies\(^1\).

Next we conducted keyword searches in ERIC, PsycINFO, EconLit, and Dissertation Abstracts databases, resulting in 9,617 documents, which we refer to as reports (a particular evaluation may consist of a series of reports). Next, we manually searched the websites of policy institutes (e.g., RAND, Mathematica, NIEER) and state and federal departments (e.g., U.S. Department of Health and Human Services), as well as references mentioned in collected studies and other key early childhood education reviews. This search resulted in another 692 possible reports for inclusion in the database. In sum, 10,309 reports for possible inclusion in the early childhood

---

\(^1\) The original Abt database included ECE programs evaluated between 1960 and 2003 and used similar search techniques; therefore, we did not re-search for evaluations conducted during these years, with the exception of 2003. We conducted searches for evaluations completed between 2003 and 2007. However, our search process did result in the identification of several evaluations published prior to 2002 that were not included in the Abt database.
education portion of our database were identified, in addition to the 624 previously coded by Abt and NIEER.

Next, we developed criteria for the inclusion of studies into our meta-analytic database. In addition to being an early childhood education intervention or program from 1960 to 2007, studies had to have a treatment and control/comparison group, not simply assessing the growth of one group of children over time. Each of the groups in the study must have included at least 10 participants and incurred less than 50% attrition. Studies were excluded if they were testing a pharmacological agent, assessed children with medical disorders or learning disabilities, or tested the effectiveness of medical procedures or health-related products.

Studies must use random assignment appropriately or one of the following quasi-experimental designs: change models, fixed effects modes, regression discontinuity, difference in difference, propensity score matching, interrupted time series, instrumental variables and some other types of matching. Studies that used quasi-experimental designs must have had pre and post test information on the outcome or establish baseline equivalence of groups on several demographic characteristics determined by a joint-test. It was the goal of this meta-analysis to use more rigorous inclusion criteria than previous meta-analyses on this topic and for the quality of quasi-experimental studies that are included to be as close to approximating random assignment as possible.

After preliminary screening of abstracts of early childhood education studies, the vast majority, 91%, of the 10,309 reports were excluded due to violating our inclusion criteria. Most of the excluded reports did not meet the research design criteria, while others were eliminated for methodological errors, or did not meet our eligibility criteria. The resulting database, which is 75% completed, currently contains data from approximately 300 reports. We expect a total of about 400 reports, representing approximately 150 ECE studies for children in programs between birth and age 5 when we complete our coding this winter.

**Coding Studies.** A team of 9 graduate research assistants (4 at Harvard, 2 at Irvine and 3 at Wisconsin) were trained as coders during a 3- to 6-month process that included instruction in evaluation methods, using the coding protocol, and computing effect sizes. Trainees were paired with experienced coders in multiple rounds of practice coding. Before coding independently, research assistants also passed a reliability test comprised of randomly selected codes from a randomly selected study. In order to pass the reliability test, researchers had to calculate 100% of the effect sizes correctly and achieve 80% agreement with a master coder for the remaining codes. In instances when research assistants were just under the threshold for effect sizes, but were reliable on the remaining codes, they underwent additional effect size training before coding independently and were subject to periodic checks during their transition. Questions about coding were resolved in weekly research team conference calls involving all four principal investigators, and decisions were kept in an annotated codebook so that decisions about ambiguities could be recalled when coding subsequent studies.

**Database.** Our database consists of three levels of data: study, contrast, and effect size. Studies are defined as independent investigations of collected data. Contrasts are group comparisons within study (i.e. Head Start vs. non-Head Start, Literacy Intervention vs. no Literacy Intervention, etc…). Effect sizes are comparisons of effects between contrasts on dependent measures which include measures of cognition, achievement, behavior, socio-emotional, and
more. Studies can include multiple contrasts and sub-contrasts and multiple dependent measures. We currently have 162 studies, 882 contrasts and sub-contrasts, and 6,970 effect sizes that are non-missing in our database. We will continue to add studies into early 2011.

**Effect Size Computation.** This study’s outcome measures are reported using effect sizes as the unit of measurement. Effect sizes are computed using the Comprehensive Meta-Analysis computer software program (Borenstein, Hedges, Higgins, & Rothstein, 2005). Hedge’s G is the effect size calculation utilized by this meta-analysis. Hedges’ G is an effect size statistic that makes an adjustment to the standardized mean difference (Cohen’s d) to account for bias in the d estimator when sample sizes are small.

**Measures.** Outcome measures for this analysis cover child cognition, achievement, behavior, and socio-emotional outcomes. Cognitive outcomes include measures of theory of the mind, attention, vocabulary, IQ, task persistence, syllabic segmentation such as elision and rhyming. Achievement measures include reading, math, letter recognition, numeracy other than conservation of number, and other achievement tests. Behavior outcomes include Behavior outcomes include health risk behavior, mental health, aggressive / antisocial behavior, delinquency, internalizing, externalizing, developmental disorders, self-esteem, anxious or depressive behavior, withdrawal, impulsive or hyperactive behavior, locus of control. Socio-emotional outcomes include labeling of emotions, delay of gratification/frustration tolerance, pos/neg emotional expression, attachment, social skills, social problem solving.

The independent variables of interest for this analysis include three measures of timing: starting age, length of program, and elapsed time. Starting age is the age of the child at the beginning of the program/program. Length of the program is the amount of time (in months or years) that the program lasted. Elapsed time is the time elapsed (in months or years) after the program ended when a follow-up test occurred.

Other variables controlled for in this analysis include measures of reliability, quasi-experimental or random assignment study, attrition, whether the study controlled for baseline measures, activity level of the control group (active or passive), and whether the study was published in a peer reviewed journal.

**Data Analysis**

Following convention, we express our model in two-level (contrasts within studies and effect sizes within contrasts) hierarchical form which relates effect sizes to the child’s: i) age of entry into the program; ii) duration of program and iii) time since the completion of the program. Here the first level of the two-level model is:

\[
ES_{ijt} = \beta_0 + \beta_{1i} \text{StartAge}_{ijt} + \beta_{2i} \text{ProgDuration}_{ijt} + \beta_{3i} \text{TimeSinceProgram}_{ijt} + \beta_{11i} x_{1ijt} + \ldots + \beta_{ki} x_{kijt} + e_{ijt}
\]

where the effect size \(j\) in contrast \(i\) at measurement time \(t\), is modeled as a function of the intercept \((\beta_0)\), which represents the average (covariate adjusted) effect size for all contrasts, \(\text{StartAge}\) – the age of the child at the beginning of the program; \(\text{ProgDuration}\) – the duration of the Head Start program in years; \(\text{TimeSinceProgram}\) – the number of years between the end of the program and the outcome measurement; the \(x\)’s represent measures of program.
characteristics, child and family characteristics and study quality measures; and $e_{ijt}$ is a within-contrast error term.

The level-2 equation (contrast level) models the intercept as a function of the grand mean effect size ($\beta_0$) and a between-contrast random error term ($u_i$):

$$\beta_{0i} = \beta_0 + u_i$$

To facilitate interpretation of coefficients, all three key variables will be expressed in years. We will experiment with several weighting schemes to take into consideration the within-study variance, within-study dependency, between-study variance, and the sampling error. Specifically, non-iterative and iterative method of moments and weighted hierarchical linear models will be conducted to generate the weighting matrix (Hedges, Tipton, & Johnson, 2010; Stevens & Taylor, 2009; Raudenbush & Bryk, 1985). Since it is unlikely that the effects of each of these variables would be linear, we will experiment with a variety of theoretically-appropriate nonlinear forms. For example, persistence will be estimated using a negative exponential and in more flexible ways using dummy variables. Moreover, we will not assume that the same functional form will fit all outcomes, given, for example, evidence from both Perry and Abecedarian of longer-lived program effects on achievement than IQ. We will also test for such theoretically appropriate interactions such as program duration by time since completion of program to assess whether the longest program have the most enduring impacts.

**Findings / Results:**
*Description of the main findings with specific details.*

Given the incomplete nature of our data base, our preliminary data are intended only to provide a rough idea of our likely results. The first two columns of Table 1 provide descriptive statistics on our key timing variables and on other measures we intend to include in our model. Our descriptive show that the average starting age of our programs is at 3.8 years. This will fall as additional studies are added from 0-3 age range. The average length of program is approximately one year and the average follow-up time after treatment is approximately two years. The final column of Table 1 shows regression coefficients and standard errors from a very preliminary model of our timing measures. Significant negative effects are estimated for post-treatment time, which suggests that treatment effects tend to be the highest immediately following the end of treatment. In this model, length of program has an unexpected negative sign, although both this coefficient and the one on starting age are not statistically significant. These results are likely to change as we continue to add data and conduct robustness checks.

**Conclusions:**
*Description of conclusions, recommendations, and limitations based on findings.*

Given the preliminary nature of our analysis, we cannot offer conclusions at this point.
Appendices
Not included in page count.

Appendix A. References
References are to be in APA version 6 format.


Appendix B. Tables and Figures
Not included in page count.

Table 1. Preliminary Descriptive Statistics for Starting Time, Duration and Time Since Program

<table>
<thead>
<tr>
<th></th>
<th>M</th>
<th>SD</th>
<th>Regression Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Starting age (yrs)</td>
<td>3.82</td>
<td>1.20</td>
<td>-0.014 (0.025)</td>
</tr>
<tr>
<td>Program duration (yrs)</td>
<td>1.04</td>
<td>1.02</td>
<td>-0.044 (0.035)</td>
</tr>
<tr>
<td>Time since program (yrs)</td>
<td>1.95</td>
<td>5.43</td>
<td>-.013* (.006)</td>
</tr>
<tr>
<td>Number of Effect Sizes</td>
<td>---</td>
<td>---</td>
<td>3609</td>
</tr>
<tr>
<td>Number of Contrasts</td>
<td>---</td>
<td>---</td>
<td>300</td>
</tr>
<tr>
<td>Number of Studies</td>
<td>---</td>
<td>---</td>
<td>138</td>
</tr>
</tbody>
</table>

Based on partial data. Robust standard errors are given in parentheses. Models are weighted by the inverse variance multiplied by the number of effect sizes within a contrast. Dummy variables included for 4 of 5 domains of interest. Controls for other study design characteristics are also included in the regression model.

*p < 0.05, **p < 0.01, ***p < 0.001
Title: School Climate in Urban Elementary Schools: Its Role in Predicting Low-Income Children’s Transition from Early Educational RCT to Kindergarten

Author(s): Amy E. Lowenstein, C. Cybele Raver, Stephanie M. Jones, Fuhua Zhai, and Rachel A. Pess
Abstract Body
Limit 5 pages single spaced.

Background / Context:
Description of prior research and its intellectual context.

Past research on school-level factors that predict children's development has focused largely on associations between a limited number of characteristics, such as school size and school resources, and children’s academic achievement (e.g., Greenwald, Hedges, & Laine, 1996; Lee & Smith, 1997). Few studies take a more comprehensive look at the measurement of school climate or examine its relationship to children’s social-emotional competence. Studies that aim to link features of schools with student outcomes typically necessitate a multilevel approach because students are nested in schools (Rumberger & Palardy, 2004). There are numerous methodological challenges involved in studying school effects on student outcomes, including the threats of selection bias and model misspecification, the need for a large sample size (especially at the school level), and the reliable measurement of school-level constructs of interest. Under ideal circumstances, studies of school effectiveness involve experimental or quasi-experimental designs that allow for causal inference. However, in the case of long-term follow-ups of children enrolled in randomized controlled trials (RCTs), education scientists typically do not have the luxury of random assignment to post-intervention treatments.

In other analyses based on the current study, the Chicago School Readiness Project (CSRP), we use propensity score matching, a quasi-experimental approach, to examine long-term effects of participation in a preschool intervention program on children's academic and behavioral outcomes after accounting for the quality of schools that they attended subsequently (Zhai & Raver, 2010). The benefit of methods such as these is that they allow for stronger internal validity. However, they do so to the exclusion of an in-depth characterization of the constructs at hand. In contrast, structural equation modeling (SEM) allows for the nuanced but parsimonious description of constructs of interest, such as school quality, using a latent-variable framework. The emphasis in SEM is on measurement, sometimes at the expense of internal validity.

A substantial body of literature suggests that low-income children are at greater risk for emotional and behavioral problems than their higher-income peers (Duncan, Brooks-Gunn, & Klebanov, 1994). Children spend most of their waking hours in school, yet little is known about whether and how school contexts support or constrain low-income children’s social-emotional development. Furthermore, most research on school climate has focused on middle- and high-school students. Far less is known about the role of school climate in the development of elementary-school children (Rumberger & Palardy, 2004). Finally, the research base on schools and children’s social-emotional development is quite thin; most research in this area has focused on achievement outcomes. Unlike many other studies of early elementary school, our study includes reliable measures of children’s social-emotional competence. In the current paper, we capitalize on these strengths and the measurement capabilities of SEM to develop a new, multidimensional model of school climate, which we use to predict low-income children’s social-emotional development during the transition to kindergarten. We use follow-up data from a cluster-randomized controlled trial of a classroom-based intervention in Head Start classrooms.
Purpose / Objective / Research Question / Focus of Study:
Description of the focus of the research.

The purposes of this study were to: (1) Identify a multidimensional model of school climate and (2) use it to predict low-income children’s social-emotional outcomes during the transition to kindergarten.

Setting:
Description of the research location.

The research setting consisted of kindergarten classrooms located in the Chicago public schools (CPS).

Population / Participants / Subjects:
Description of the participants in the study: who, how many, key features or characteristics.

The original CSRP sample consisted of 602 children enrolled in 35 classrooms nested within 18 Head Start sites that were randomly assigned to either an intervention program designed to support young children’s self-regulation or to a control group (see Raver et al., 2009). Enrolled children initially ranged from 3 to 5 years of age.

The sample used in the current analysis consisted of 200 children nested in kindergarten classrooms in 82 public schools. Children ranged from 4 to 6 years of age. Twenty percent were Hispanic and 74% were African American (see Table 1). The vast majority were from low-income households.

Intervention / Program / Practice:
Description of the intervention, program or practice, including details of administration and duration.

Children were enrolled in CPS schools in fall of the kindergarten year. Most of these schools served grades pre-K to 8. School size ranged from 98 students to 5,452 students and the percentage of low-income students in the school ranged from 9% to 100% (see Table 1).

Research Design:
Description of research design (e.g., qualitative case study, quasi-experimental design, secondary analysis, analytic essay, randomized field trial).

We use multilevel factor analysis and a wide array of school-climate variables to develop a new, multidimensional model of school climate. Then, we use the resulting latent school-context factors and multilevel SEM to predict low-income children’s social-emotional development during the transition to kindergarten. We use auto-lagged models in which we control for children’s social-emotional competence in Head Start in order to account for baseline levels of social-emotional competence and capture change in children’s development between preschool and kindergarten.
Data Collection and Analysis:
Description of the methods for collecting and analyzing data.

Teachers reported on conflict and closeness in the teacher-child relationship using the Student-Teacher Relationship Scale (STRS) (Pianta, 2001) and on children’s social competence using the Social Competence and Behavior Evaluation Scale—Short Form (SCBE-30) (LaFreniere & Dumas, 1995) in spring of the Head Start year and winter of the kindergarten year. Information on school characteristics, including student demographics and the percentage of students who met or exceeded state reading and math standards, were drawn from the Illinois State Report Card and a CPS report. A survey administered to CPS students assessed student perceptions of school climate, including school safety, student support, and social-emotional learning services. Parents reported on child and family demographic characteristics.

In preliminary analyses, we used 2-level auto-lagged hierarchical linear models (HLM) to predict children’s conflict and closeness with the teacher and social competence from the following set of school characteristics: school size, student mobility rate, percentage of low-income students in the school, percentage of students in the school with an Individualized Education Plan (IEP), and percentage of students in the school who met/exceeded state reading and math standards. We included child-level controls for children’s conflict and closeness with the teacher and social competence in Head Start, as well as child and family characteristics.

Multilevel factor analysis (Muthen, 1991) will be used to identify a multidimensional model of school climate using the full set of school-level variables, and multilevel SEM will be used to predict teacher-child conflict and closeness and children’s social competence in kindergarten from these latent school factors, controlling for children’s social-emotional competence in Head Start and child and family characteristics. The reduced-form equation will be as follows:

Social-emotional functioning at kindergarten for child i in school j = B_{0ij} + B_{1ij}*(Child’s race/ethnicity) + B_{2ij}*(Child is male) + B_{3ij}*(Child’s age) + B_{4ij}*(Child’s social-emotional functioning in Head Start) + B_{5ij}*(Family income-to-needs ratio) + B_{6j}*(Latent school-climate factors) + u_{j} + e_{ij}

Findings / Results:
Description of the main findings with specific details.

Descriptive statistics revealed substantial variation in children’s social-emotional functioning and school characteristics at kindergarten (see Table 1). Preliminary results from 2-level unconditional HLM models suggest that a small but significant portion of the variance in children’s social-emotional functioning was attributable to between-school differences (ICCs ranged from 0.09 to 0.25). Additional 2-level HLM analyses in which children’s conflict with the teacher, closeness with the teacher, and social competence in kindergarten were predicted from a set of school characteristics and child-level controls indicated that a large school size was associated with a small but significant increase in teacher-child conflict ($B = 0.003, p < .05$) and a small but significant decrease in children’s social competence ($B = -0.004, p < .01$) between preschool and kindergarten (see Table 2). In contrast, a large percentage of children with an IEP was associated with a marginally significant decrease in teacher-child conflict and a marginally significant increase in children’s social competence. The inclusion of controls for children’s
social-emotional functioning in Head Start makes these models rigorous and conservatively specified, allowing for greater precision in our estimates.

**Conclusions:**
*Description of conclusions, recommendations, and limitations based on findings.*

School characteristics appear to be related to low-income children’s social-emotional competence during the transition to kindergarten. Specifically, children in larger schools show somewhat lower levels of social-emotional competence than those in smaller schools, and there is some evidence that children in schools with a larger percentage of students with an IEP show higher levels of social-emotional competence than those in schools with fewer such students. Taken together, these results suggest that a sense of community in the school and the extent to which individual students’ needs are being met matter for children’s social-emotional development.

There are several limitations to this study. First, it is correlational in nature, so we cannot draw any causal inferences about the effects of school climate on children’s development. In order to reduce the threat of selection bias, however, we controlled for children’s social-emotional competence in Head Start and child and family characteristics in all models. Second, the results presented here reflect school climate in a sample of urban elementary schools and cannot necessarily be generalized to middle schools, high schools, or elementary schools in more rural areas. Furthermore, the children in our sample were mostly racial/ethnic minorities from low-income families, so our results may not be generalizable to other children. Finally, the baseline measures of children’s social-emotional competence were collected less than a year before the outcome measures, making it harder to obtain large effects.

The presentation will include results of our school-climate factor analysis and multilevel SEM analyses in which we predict children’s social-emotional development in kindergarten from a set of school-climate factors. The discussion will highlight the importance of a multidimensional approach to the measurement of school climate.
Appendices
Not included in page count.

Appendix A. References
References are to be in APA version 6 format.

### Table 1

**Descriptive Statistics for Variables Used in HLM Analyses**

<table>
<thead>
<tr>
<th>Variable</th>
<th>M</th>
<th>SD</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Outcome variables (all measured at kindergarten)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conflict with the teacher (STRS composite)</td>
<td>12.62</td>
<td>6.81</td>
<td>7-33</td>
</tr>
<tr>
<td>Closeness with the teacher (STRS composite)</td>
<td>32.96</td>
<td>5.77</td>
<td>16-40</td>
</tr>
<tr>
<td>Social competence (SCBE composite)</td>
<td>29.34</td>
<td>12.83</td>
<td>0-50</td>
</tr>
<tr>
<td><strong>Child and family characteristics (n = 200)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child’s race/ethnicity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White/non-Hispanic</td>
<td>0.04</td>
<td>0.19</td>
<td>0-1</td>
</tr>
<tr>
<td>African American</td>
<td>0.74</td>
<td>0.44</td>
<td>0-1</td>
</tr>
<tr>
<td>Hispanic</td>
<td>0.20</td>
<td>0.40</td>
<td>0-1</td>
</tr>
<tr>
<td>Bi-racial</td>
<td>0.03</td>
<td>0.16</td>
<td>0-1</td>
</tr>
<tr>
<td>Child is male</td>
<td>0.52</td>
<td>0.50</td>
<td>0-1</td>
</tr>
<tr>
<td>Child’s age at kindergarten (years)</td>
<td>4.61</td>
<td>0.67</td>
<td>4-6</td>
</tr>
<tr>
<td>Conflict with the teacher (STRS composite) in Head Start</td>
<td>11.73</td>
<td>5.17</td>
<td>7-28</td>
</tr>
<tr>
<td>Closeness with the teacher (STRS composite) in Head Start</td>
<td>33.15</td>
<td>5.92</td>
<td>12-40</td>
</tr>
<tr>
<td>Social competence (SCBE composite) in Head Start</td>
<td>29.46</td>
<td>9.84</td>
<td>1-49</td>
</tr>
<tr>
<td>Family income-to-needs ratio at kindergarten</td>
<td>0.69</td>
<td>0.58</td>
<td>0-3.45</td>
</tr>
<tr>
<td><strong>School characteristics (all measured at kindergarten; n = 82)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>School size (total enrollment)</td>
<td>765.11</td>
<td>641.27</td>
<td>98-5,452</td>
</tr>
<tr>
<td>Student mobility rate</td>
<td>24.83</td>
<td>15.36</td>
<td>2.2-84.4</td>
</tr>
<tr>
<td>Percentage of low-income students in the school</td>
<td>90.38</td>
<td>14.43</td>
<td>8.9-100</td>
</tr>
<tr>
<td>Percentage of students in the school with an Individualized Education Plan (IEP)</td>
<td>12.36</td>
<td>4.30</td>
<td>5-30.6</td>
</tr>
<tr>
<td>Percentage of students in the school who met/exceeded state reading and math standards</td>
<td>59.73</td>
<td>15.72</td>
<td>32.8-97.7</td>
</tr>
<tr>
<td>Variable</td>
<td>Conflict with the teacher</td>
<td>Closeness with the teacher</td>
<td>Social competence</td>
</tr>
<tr>
<td>----------</td>
<td>--------------------------</td>
<td>---------------------------</td>
<td>------------------</td>
</tr>
<tr>
<td><strong>Child and family characteristics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child’s race/ethnicity (Hispanic is omitted category)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White/non-Hispanic</td>
<td>0.52 (1.56)</td>
<td>2.43† (1.26)</td>
<td>0.68 (4.42)</td>
</tr>
<tr>
<td>African American</td>
<td>3.13* (1.36)</td>
<td>-0.81 (1.66)</td>
<td>-2.65 (3.21)</td>
</tr>
<tr>
<td>Bi-racial</td>
<td>-6.65* (2.80)</td>
<td>-1.79 (2.13)</td>
<td>2.73 (3.07)</td>
</tr>
<tr>
<td>Child is male</td>
<td>0.36 (0.92)</td>
<td>-1.86† (1.11)</td>
<td>-3.42 (2.66)</td>
</tr>
<tr>
<td>Child’s age at kindergarten (years)</td>
<td>-0.81 (0.76)</td>
<td>-0.24 (0.85)</td>
<td>1.49 (1.63)</td>
</tr>
<tr>
<td>Conflict with the teacher in Head Start</td>
<td>0.45*** (0.09)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Closeness with the teacher in Head Start</td>
<td></td>
<td>0.16* (0.08)</td>
<td></td>
</tr>
<tr>
<td>Social competence in Head Start</td>
<td></td>
<td></td>
<td>0.19* (0.10)</td>
</tr>
<tr>
<td><strong>Family income-to-needs ratio at kindergarten</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-1.42 (0.88)</td>
<td>1.48† (0.82)</td>
<td>3.17† (1.65)</td>
</tr>
<tr>
<td><strong>School characteristics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>School size (total enrollment)</td>
<td>0.00* (0.00)</td>
<td>-0.00 (0.00)</td>
<td>-0.00** (0.00)</td>
</tr>
<tr>
<td>Student mobility rate</td>
<td>0.02 (0.04)</td>
<td>0.00 (0.05)</td>
<td>0.08 (0.10)</td>
</tr>
<tr>
<td>% of low-income students in the school</td>
<td>-0.06 (0.05)</td>
<td>-0.01 (0.05)</td>
<td>-0.02 (0.10)</td>
</tr>
<tr>
<td>% of students in the school with an Individualized Education Plan (IEP)</td>
<td>-0.22† (0.13)</td>
<td>0.08 (0.15)</td>
<td>0.53† (0.30)</td>
</tr>
<tr>
<td>% of students in the school who met/exceeded state reading and math standards</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.00 (0.06)</td>
<td>-0.01 (0.04)</td>
<td>0.01 (0.10)</td>
</tr>
<tr>
<td><strong>Constant</strong></td>
<td>16.17† (8.60)</td>
<td>29.94** (8.55)</td>
<td>13.08 (16.23)</td>
</tr>
</tbody>
</table>

*Note. Level-1 equation: Social-emotional functioning at kindergarten = B0 + B1*(Child’s race/ethnicity) + B2*(Child is male) + B3*(Child’s age) + B4*(Child’s social-emotional functioning in Head Start) + B5*(Family income-to-needs ratio) + R. Level-2 equation: B0 = G00 + G01*(School size) + G02*(Student mobility rate) + G03*(% low-income students in school) + G04*(% students in school with IEP) + G05*(% students in school who met/exceeded state reading/math standards) + U0.

† p < .10. * p < .05. ** p < .01. *** p < .001.
Title: Complex Setting-Level Interventions in Social-Emotional Learning: Testing Multi-Level Causal Processes in the School-Randomized Evaluation of the 4Rs Program.

Author(s):
Stephanie M. Jones, Ph.D.
Assistant Professor
Harvard Graduate School of Education
14 Apprian Way, Larsen 603
Cambridge, MA 02138
617-496-2223 phone
email: jonesst@gse.harvard.edu

Joshua L. Brown, Ph.D.
Assistant Professor
Department of Psychology
Fordham University
441 East Fordham Road
Bronx, NY 10458-9993
718-817-4069 phone
718-817-3785 fax
email: cjobrown@fordham.edu

J. Lawrence Aber, Ph.D.
Department of Applied Psychology
Steinhardt School of Education
New York University
Kimball Hall, Room 417E
246 Greene Street
New York, N.Y. 10003
tel: 212-998-5410
fax: 212-995-4215
email: lawrence.aber@nyu.edu
Abstract Body
Limit 5 pages single spaced.

Background/context:
Description of prior research, its intellectual context and its policy context.
Over the last two decades, developmental science has made significant progress in understanding children’s trajectories of social-emotional and academic development (e.g., Duncan et al., 2007; Miles & Stipek, 2006). At the same time, there has been substantial growth in the design, implementation, and rigorous evaluation of school-based interventions to promote positive social-emotional development and/or academic achievement (e.g., IES, 2003; Embry, 2002; Kellam et al., 2008). What distinguishes this new generation of preventive intervention evaluation studies is their increasingly sophisticated and fine-grained multi-level theories of change, in which processes at multiple ecological levels are targeted (e.g., individuals and social settings such as classrooms). Such advances in the use of developmental-ecological theory, epidemiological data, and small-scale, highly controlled efficacy studies to articulate a coherent multi-level logic model are at the core of prevention science (e.g., Kellam & Langevin, 2003). Moreover, theory- and data-driven logic models have been useful guides to the development of targeted measurement strategies, to refined expectations for change in proximal and distal outcomes, and to a growing literature on the science of replicating and sustaining programs beyond the original demonstration. Indeed, using such models as a guide to unpacking the black box of intervention effects by testing process mediators is a critical step for identifying the key features of programs that are likely to promote successful replication and sustainability, as well as for supporting program improvement efforts (Coie et al., 1993). To date however, too little attention has been paid to the methodological challenges associated with drawing causal inferences from experimental evaluations of programs that hold multi-level theories of change.

Purpose / objective / research question / focus of study:
Description of what the research focused on and why.
The present study employs, and then compares, two methodological approaches to statistical mediation to examine the mediating role of classroom quality in the experimental impact of the 4Rs Program (Reading, Writing, Respect and Resolution) on children’s social-cognitive and behavioral development (Jones, Brown & Aber, 2008). Specifically, this paper builds upon established main effects of the 4Rs Program (e.g., Jones, Brown & Aber, in press; Brown, Jones, LaRusso & Aber, 2010) to examine the following broad question: Do intervention-induced changes in the emotional and instructional climate of the classroom as rated at the classroom-level by independent observers underlie intervention-induced changes in children’s socially competent and aggressive behavior, and their attendance and achievement? Thus, in this paper we conduct an analysis of a group-randomized educational intervention in which the treatment is randomized at the school level, the mediator is at the classroom level, and the outcome is at the child level. Findings from this work have implications both for basic theory in developmental science and for the continued evolution of school-based strategies to promote children’s social-emotional learning and thereby prevent some of the most ubiquitous mental health problems of middle childhood.

Setting:
Description of where the research took place.
This project took place over three consecutive school years in 18 public elementary schools in New York City (in four of the five boroughs).

Population / Participants / Subjects:
Description of participants in the study: who (or what) how many, key features (or characteristics).
Participants were 1184 children (49% boys; average age at time 1 = 8.17 yrs, SD = 0.7), and 146 classrooms (88% female; average age = 35) in 18 public inner-city elementary schools in a large metropolitan city in the Eastern United States. The children and their teachers/classrooms are part of an ongoing, longitudinal evaluation of a universal, school-wide literacy and social-emotional learning prevention program (4Rs: Reading, Writing, Respect and Resolution) implemented for three consecutive years in 9 intervention (n = 630; 53.2%) and 9 control (n = 554; 46.8%) schools. Data were gathered from children, their teachers, and in their classrooms over six longitudinal waves across 3 school years. Because the 4Rs Program was randomized at the school-level, children who moved out of a participating school were not followed (e.g., 58 children in Wave 2) and consent was requested for new children who moved into a participating school at each follow-up wave (e.g., 124 children at Wave 2. Attrition between waves was minimal (on average 8.6%) and was primarily due to student mobility out of participating schools. Refusals from parents to continue were rare (e.g., n = 1 in Wave 2).

According to parent-reports at baseline, 52% (n = 425) of children lived in a single-parent household, 14.5% (n = 119) of parents were unemployed, 29.7% (n = 243) of parents had less than a high school diploma or GED, and 60.9% (n = 498) of households were at or below 100% of the federal poverty level. Based on parent-reports at baseline and NYC Department of Education records when parent-reports were missing, children represented diverse racial/ethnic groups; 45.3% (n = 368) were Hispanic/Latino, 41.1% (n = 334) Black/African American, 5% (n = 41) non-Hispanic White, and 8.6% (n = 70) represented other racial/ethnic groups (e.g., Asian, Pacific Islander, Native American).

Intervention / Program / Practice:
Description of the intervention, program or practice, including details of administration and duration.
The 4Rs Program (Reading, Writing, Respect and Resolution) is a universal, school-based intervention in literacy development and social-emotional learning that integrates a focus on social and emotional development into the language arts curriculum for children in grades K-5. In this evaluation, Developed and run by a community-based non-profit organization called the Morningside Center for Teaching Social Responsibility, the 4Rs Program uses high quality children’s literature as a springboard for helping students gain skills and understanding in the areas of handling anger, listening, assertiveness, cooperation, negotiation, mediation, building community, celebrating differences, and countering bias. By highlighting universal themes of conflict, feelings, relationships, and community, the 4Rs curriculum adds social and emotional meaning and skill building to rigorous literacy instruction. The 4Rs Program has two primary components: (1) a comprehensive 7-unit, 21-35 lesson, literacy-based curriculum in social-emotional learning and (2) 25 hours of training followed by ongoing coaching of teachers to support them in teaching the 4Rs curriculum with a minimum of 12 contacts in one school year.

Research Design:
Description of research design (e.g., qualitative case study, quasi-experimental design, secondary analysis, analytic essay, randomized field trial).
Forty-one schools representative of the population of NYC elementary schools were originally identified as potential participants in the 4Rs evaluation. Of these 41, 24 agreed to the process of matching and randomization. Prior to randomization, a pairwise matching procedure was used to maximize demographic similarity of intervention and control groups. An algorithm was used to compute the distance from each school to every other school along 20 demographic and school characteristics. These variables were drawn primarily from the 2001-2002 administrative databases kept by the city’s Department of Education and were selected to represent a number of important dimensions related to the criterial outcomes. To conduct random assignment of matched pairs to 4Rs intervention and control groups, a MatLab uniform random numbers generator was employed to generate, in sequence, 12 random numbers ranging from 0-1 that were assigned to the first school in each of the 12 pairs (note, a total of 24 schools were recruited to participate in this study and were matched into 12 pairs, the 9 best matching pairs were kept as study schools and 3 pairs were kept as back-ups). The first school in each pair was assigned to the intervention or control group based on the randomly generated number, and the second school in the pair was, therefore, assigned to the other group. After random assignment, the two groups were compared across the 20 demographic characteristics employed in the matching procedures. As expected the two groups did not differ significantly on any of these characteristics and eta² values (the proportion of variance in the demographic characteristic explained by differences between the two groups) were minimal. Based on these statistics, the schools can be described as racially and ethnically diverse, composed primarily of students who receive a free school lunch, and characterized by attendance rates over 89% and one-year stability rates that range from 86% to 95%.

**Data Collection and Analysis:**

*Description of the methods for collecting and analyzing data.*

Consent packages (in English and Spanish) were sent home to all parents of third grade children in the 18 participating schools informing them of the study and seeking consent for their child to participate. The overall consent rate was 64.54% across schools (range = 44% to 79%); consent rates did not differ between treatment (65.2%) and control (63.7%) schools. Non-participants included children whose parents did not speak English or Spanish well enough to consent to participate and special needs children who could not be interviewed even on an individual basis (e.g., due to autism).

**Child Outcomes:** At each wave teachers completed questionnaires rating the language and literacy skills, as well as social competence and externalizing problems of each child in their class with consent to participate. Teachers were paid at the union wage of $36.50/per hour for completion of the surveys at each assessment. At each wave, children also completed questionnaires rating their aggressive social-cognitions, pro-social-cognitions, and internalizing symptoms. Data were collected from the children in small class groups (n = 5 to 20). All questions were read out loud by a research assistant while a second research assistant circulated to monitor children’s placement of responses and to answer the children’s questions. Children who did not have consent to participate or who refused assent worked on an alternative activity with their classroom teacher.

**Classroom Mediators:** Classroom quality was assessed via independent observation using the Classroom Assessment Scoring System (Pianta, LaParo, & Hamre, 2008) in 3rd grade classrooms during the spring of the first year of the study (n=82), in 3rd and 4th grade classrooms during the
fall and spring of the second year (n=147), and in 3rd, 4th and 5th grade classrooms during the fall and spring of the third year (n=191). Observations were conducted by a multiracial/multiethnic field research team who received extensive training to reliability standards in the use of this instrument. Classrooms were observed for two hours, divided into four 20-minute periods of observation, each observation period followed by a 10-minute rating period. CLASS assesses three primary domains of classroom climate in preschool through fifth grade classrooms: Emotional Support, Classroom Organization, and Instructional Support. Each broad domain is comprised of several specific dimensions of interactions. Each dimension, in turn, is represented by a continuum of indicators of that dimension, each indicator including a behaviorally anchored, observable description of interactions in the classroom (teacher-student, student-student). Internal reliability was .90 for Emotional Support, .83 for Classroom Organization, and .90 for Instructional Support and .93 for a composite index of all three subscales. There is strong support for the psychometric properties of the CLASS, including demonstrated relationships to children’s social and academic development (predictive validity) both during the preschool (e.g., Howes et al., 2008) and elementary school years (e.g., Pianta et al., 2008).

Mediation will be addressed using two primary methods. To begin, more traditional causal step approaches (Baron & Kenny, 1986) to mediation will be employed. We will also employ instrumental variables techniques, in which random assignment of units is employed as an instrument to generate an unbiased test of the causal effects of a key mediator on outcomes. Instrumental variable analysis in this case proceeds in two steps, the first estimating the relationship between intervention status and the key mediator (e.g., random assignment to school-based SEL intervention or not and classroom quality), and the second, estimating the relationship between the intervention-predicted value of the mediator (classroom quality) and children’s outcomes (e.g., behavior problems).

Findings / Results:
Description of main findings with specific details
Findings to date through the second year of the intervention are summarized below. Findings are presented below by broad outcome domain.
Classroom Quality. Brown et al. (2010) report that after one year of intervention, classrooms in 4Rs schools were rated by independent observers as higher in overall quality compared to classrooms in control schools. Specifically, this work indicated that 4Rs classrooms had higher mean levels of classroom emotional support and instructional support (but not classroom organization) than control classrooms at the end of the first year of the intervention.
Social Cognitive Processes and Social-Emotional Symptomatology. Over two consecutive years, children in 4Rs schools self-reported slower rates of increase in hostile attributional bias, a slowed rate of growth in aggressive interpersonal negotiation strategies that appears to begin toward the outset of the second year of exposure to intervention, and a steeper rate of decline in depressive and ADHD symptoms compared to children in the control schools.
Aggressive and Socially Competent Behavior. Teachers in 4Rs schools reported slower growth in children’s aggressive behavior (compared to increases in control schools), and increases in social competence (compared to declines in control schools) over two school years.
Academic Functioning. While there were no main effects of treatment on teacher reports of children’s academic skills or on the three school records outcomes examined, there were treatment by baseline behavioral risk interactions for standardized math and reading achievement and for teacher reported academic skills. In short, children identified by teachers at greatest
behavioral risk at baseline showed greater improvements as a result of exposure to 4Rs in their math and reading achievement and in teacher reports of their academic skills (see Figure 1 in Appendix B). Importantly, this set of treatment by baseline behavioral risk interactions were not evident for the social-emotional outcomes, regardless of the type of model examined (i.e., as a growth model estimating treatment effects on growth parameters, or a basic point-in-time model estimating treatment effects on Wave 4, controlling for baseline levels). This suggests that these treatment by risk interactions are not an artifact of the form of model applied, but instead are tied to the developmental domain examined: children’s academic functioning and not, in this case, their social-emotional skills.

Over all, the treatment main effects reported after two years are small to moderate in size (ranging in size from .05 for the teacher-reported aggression slope to .22 for the depressive symptoms slope). In contrast, the interactions of treatment with baseline behavioral risk are represented by treatment effects for the highest risk group of moderate size (ranging from .56 for math achievement and academic skills to .60 for reading achievement).

**Conclusions:**

*Description of conclusions and recommendations based on findings and overall study.*

Our findings to date on the impacts of an integrated, social-emotional and literacy program provides clear evidence that this universal intervention has both broad impacts on classroom quality, child-level social-cognitive processes and behaviors in the social-emotional domain, and targeted impacts in the academic domain. This study provides good evidence that universal school-based interventions, delivered to whole populations of children, can result in substantial impacts on children’s developmental health and well-being and well as on classroom settings. The present study is designed to begin to address questions of process, focusing directly on whether changes in classroom characteristics underlie changes in children’s developmental outcomes.
Appendices
Not included in page count.

Appendix A. References
References are to be in APA version 6 format.
Appendix B. Tables and Figures
Not included in page count.

Figure 1. Interaction of treatment and baseline behavioral risk on Year 2 math achievement

![Interaction of treatment and baseline behavioral risk on Year 2 math achievement](image-url)